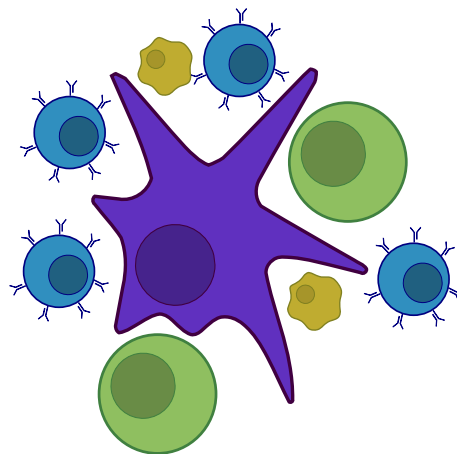


Computational Solutions for Addressing Heterogeneity in DNA Methylation Data

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes



von

Michael Scherer

Saarbrücken

2020

Tag des Kolloquiums: 30.03.2021

Dekan: Univ.-Prof. Dr. Thomas Schuster

Prüfungsausschuss

Vorsitz: Prof. Dr. Olga Kalinina

Berichterstatter/Gutachter: Prof. Dr. Dr. Thomas Lengauer

Prof. Dr. Jörn Walter

Prof. Dr. Tobias Marschall

Akad. Mitarbeiter: Dr. Gilles Gasparoni

*"While we live according to race, colour or creed
While we rule by blind madness and pure greed
Our lives dictated by tradition, superstition, false religion
Through the eons and on and on*

*Oh yes we'll keep on tryin'
We'll tread that fine line
Oh oh we'll keep on tryin'
Till the end of time"*

Queen: Innuendo

Abstract

DNA methylation, a reversible epigenetic modification, has been implicated with various biological processes including gene regulation. Due to the multitude of datasets available, it is a premier candidate for computational tool development, especially for investigating heterogeneity within and across samples. We differentiate between three levels of heterogeneity in DNA methylation data: between-group, between-sample, and within-sample heterogeneity. Here, we separately address these three levels and present new computational approaches to quantify and systematically investigate heterogeneity.

Epigenome-wide association studies relate a DNA methylation aberration to a phenotype and therefore address between-group heterogeneity. To facilitate such studies, which necessarily include data processing, exploratory data analysis, and differential analysis of DNA methylation, we extended the R-package *RnBeads*. We implemented novel methods for calculating the epigenetic age of individuals, novel imputation methods, and differential variability analysis. A use-case of the new features is presented using samples from Ewing sarcoma patients. As an important driver of epigenetic differences between phenotypes, we systematically investigated associations between donor genotypes and DNA methylation states in methylation quantitative trait loci (methQTL). To that end, we developed a novel computational framework –MAGAR– for determining statistically significant associations between genetic and epigenetic variations. We applied the new pipeline to samples obtained from sorted blood cells and complex bowel tissues of healthy individuals and found that tissue-specific and common methQTLs have distinct genomic locations and biological properties.

To investigate cell-type-specific DNA methylation profiles, which are the main drivers of within-group heterogeneity, computational deconvolution methods can be used to dissect DNA methylation patterns into latent methylation components. Deconvolution methods require profiles of high technical quality and the identified components need to be biologically interpreted. We developed a computational pipeline to perform deconvolution of complex DNA methylation data, which implements crucial data processing steps and facilitates result interpretation. We applied the protocol to lung adenocarcinoma samples and found indications of tumor infiltration by immune cells and associations of the detected components with patient survival.

Within-sample heterogeneity (WSH), i.e., heterogeneous DNA methylation patterns at a genomic locus within a biological sample, is often neglected in epigenomic studies. We present the first systematic benchmark of scores quantifying WSH genome-wide using simulated and experimental data. Additionally, we created two novel scores that quantify DNA methylation heterogeneity at single CpG resolution with improved robustness toward technical biases. WSH scores describe different types of WSH in simulated data, quantify differential heterogeneity, and serve as a reliable estimator of tumor purity.

Due to the broad availability of DNA methylation data, the levels of heterogeneity in DNA methylation data can be comprehensively investigated. We contribute novel computational frameworks for analyzing DNA methylation data with respect to different levels of heterogeneity. We envision that this toolbox will be indispensable for understanding the functional implications of DNA methylation patterns in health and disease.

Kurzfassung

DNA Methylierung ist eine reversible, epigenetische Modifikation, die mit verschiedenen biologischen Prozessen wie beispielsweise der Genregulation in Verbindung steht. Eine Vielzahl von DNA Methylierungsdatensätzen bildet die perfekte Grundlage zur Entwicklung von Softwareanwendungen, insbesondere um Heterogenität innerhalb und zwischen Proben zu beschreiben. Wir unterscheiden drei Ebenen von Heterogenität in DNA Methylierungsdaten: zwischen Gruppen, zwischen Proben und innerhalb einer Probe. Hier betrachten wir die drei Ebenen von Heterogenität in DNA Methylierungsdaten unabhängig voneinander und präsentieren neue Ansätze um die Heterogenität zu beschreiben und zu quantifizieren.

Epigenomweite Assoziationsstudien verknüpfen eine DNA Methylierungsveränderung mit einem Phänotypen und beschreiben Heterogenität zwischen Gruppen. Um solche Studien, welche Datenprozessierung, sowie exploratorische und differentielle Datenanalyse beinhalten, zu vereinfachen haben wir die R-basierte Softwareanwendung *RnBeads* erweitert. Die Erweiterungen beinhalten neue Methoden, um das epigenetische Alter vorherzusagen, neue Schätzungsmethoden für fehlende Datenpunkte und eine differentielle Variabilitätsanalyse. Die Analyse von Ewing-Sarkom Patientendaten wurde als Anwendungsbeispiel für die neu entwickelten Methoden gewählt. Wir untersuchten Assoziationen zwischen Genotypen und DNA Methylierung von einzelnen CpGs, um sogenannte methylation quantitative trait loci (methQTL) zu definieren. Diese stellen einen wichtiger Faktor dar, der epigenetische Unterschiede zwischen Gruppen induziert. Hierzu entwickelten wir ein neues Softwarepaket (*MAGAR*), um statistisch signifikante Assoziationen zwischen genetischer und epigenetischer Variation zu identifizieren. Wir wendeten diese Pipeline auf Blutzelltypen und komplexe Biopsien von gesunden Individuen an und konnten gemeinsame und gewebespezifische methQTLs in verschiedenen Bereichen des Genoms lokalisieren, die mit unterschiedlichen biologischen Eigenschaften verknüpft sind.

Die Hauptursache für Heterogenität innerhalb einer Gruppe sind zelltypspezifische DNA Methylierungsmuster. Um diese genauer zu untersuchen kann Dekonvolutionssoftware die DNA Methylierungsmatrix in unabhängige Variationskomponenten zerlegen. Dekonvolutionsmethoden auf Basis von DNA Methylierung benötigen technisch hochwertige Profile und die identifizierten Komponenten müssen biologisch interpretiert werden. In dieser Arbeit entwickelten wir eine computerbasierte Pipeline zur Durchführung von Dekonvolutionsexperimenten, welche die Datenprozessierung und Interpretation der Resultate beinhaltet. Wir wendeten das entwickelte Protokoll auf Lungenadenokarzinome an und fanden Anzeichen für eine Tumordinfiltration durch Immunzellen, sowie Verbindungen zum Überleben der Patienten.

Heterogenität innerhalb einer Probe (within-sample heterogeneity, WSH), d.h. heterogene Methylierungsmuster innerhalb einer Probe an einer genomischen Position, wird in epigenomischen Studien meist vernachlässigt. Wir präsentieren den ersten Vergleich verschiedener, genomweiter WSH Maße auf simulierten und experimentellen Daten. Zusätzlich entwickelten wir zwei neue Maße um WSH für einzelne CpGs zu berechnen, welche eine verbesserte Robustheit gegenüber technischen Faktoren aufweisen. WSH Maße beschreiben verschiedene Arten von WSH, quantifizieren differentielle Heterogenität und sagen Tumorreinheit vorher.

Aufgrund der breiten Verfügbarkeit von DNA Methylierungsdaten können die Ebenen der Heterogenität ganzheitlich beschrieben werden. In dieser Arbeit präsentieren wir neue Softwarelösungen zur Analyse von DNA Methylierungsdaten in Bezug auf die verschiedenen Ebenen der Heterogenität. Wir sind davon überzeugt, dass die vorgestellten Softwarewerkzeuge unverzichtbar für das Verständnis von DNA Methylierung im kranken und gesunden Stadium sein werden.

Acknowledgments

I had the pleasure to work with many great scientists, and I would like to thank everyone that contributed in any way to this work. First and foremost, I would like to thank my two doctoral supervisors Thomas Lengauer and Jörn Walter. Thomas is a role model of scientific excellence and has great mentoring skills. Thank you, Thomas, for being an idol throughout my studies and for giving me the opportunity to work in your group. The way you communicate complex scientific and mathematical problems is extraordinary, and helped me a lot in improving my own teaching skills. I would also like to thank Jörn for countless scientific discussions on various biological and bioinformatic problems. The way you address computational challenges in biological systems is exceptional, and it has been a pleasure to work on new ideas for addressing these questions. You are a role model for establishing scientific collaborations and, when faced with a new problem, always know an expert in the field to answer the questions.

Next, I would like to acknowledge the people with whom I worked very closely during the projects that I will present in this work: Fabian Müller, Markus List, and Pavlo Lutsik. Fabian was not only my Master thesis advisor, he was also the person introducing me into the field of computational epigenomics. Together with Markus, he supervised a large part of my work. It has been a pleasure to work with the both of you and I learned a lot from you about the way science works. Also thanks to Pavlo for a great collaboration and for plenty of nice discussions.

In the first two years of my PhD, I worked at the department of Computational Biology and Applied Algorithmics at the Max-Planck-Institute for Informatics together with many people, and I am grateful to all of them for the discussions and entertaining lunch breaks. Most notably, I would like to mention the people from the Coffee Club: Matthias, Fabian, Markus, Lisa, Nora, Anna, Peter, Florian, Dilip, Siva, Fatemeh. The computational epigenetics group within our department provided a great atmosphere to learn about epigenetic regulation beyond DNA methylation, thanks to: Felipe, Peter, Fabian, and Christoph.

After moving to the upper quarter of the Campus to work at the Department of Genetics/Epigenetics, I had the pleasure to work in a heterogeneous group of wet lab scientists and bioinformaticians. The group seminars have always been enlightening and I thank everyone for the numerous discussions. Special thanks go to my fellow bioinformaticians Abdul and Karl. Additional thanks to Gilles and Sascha for the scientific and non-scientific discussions.

I would also like to thank all the people that contributed administrative work: Ruth, Nicole, Achim, Georg, Sascha, the HR departments of Saarland University and the MPI, and the IST at the MPI. In the last months of my PhD studies, I was financially supported by the Graduate School of Computer Science at Saarland University, and I would especially like to thank Michelle Carnell for her support.

My deepest gratitude goes to my parents for their continuous support throughout my studies, and to my whole family for being the anchor in my life. Also a huge thanks to all of my friends for their support.

I would like to thank the people proof-reading this thesis: Laurena, Sascha, Abdul, Markus, and Fabian. Last, many thanks to Tobias Marschall for agreeing to review the work and to the whole thesis committee.

Contents

	Page
1 Introduction	1
1.1 Epigenetic Regulation	1
1.2 Thesis Outline: Dissecting the Levels of DNA Methylation Heterogeneity	3
1.2.1 DNA Methylation Heterogeneity Between Phenotypes	3
1.2.2 DNA Methylation Heterogeneity Between Samples Sharing a Phenotype	5
1.2.3 DNA Methylation Heterogeneity Within Samples	5
2 Background	7
2.1 Mechanisms of Epigenetic Gene Regulation	7
2.1.1 Chromatin Structure	7
2.1.2 Regulatory Elements of Gene Expression	8
2.1.3 Histone Modifications	9
2.1.4 DNA Methylation	11
2.2 Influence of Genetic Variation on Epigenetic Regulation	14
2.3 DNA Methylation Aberrations and Their Association With Human Diseases	14
2.3.1 DNA Methylation Aberrations in Human Diseases	14
2.3.2 DNA Methylation Aberrations and Cancer Progression	15
2.4 Genome-Wide Mapping of DNA Methylation	15
2.4.1 Illumina Infinium BeadChip Arrays	16
2.4.2 Sequencing-Based Approaches	17
2.5 Basic Processing of DNA Methylation Data	20
2.5.1 BeadChip Arrays: From Intensity Data to a Data Matrix	20
2.5.2 Sequencing-Based Approaches: Quality Control, Alignment, Quantification	22
2.6 Methodological Background	24
2.6.1 Notations and Definitions	24
2.6.2 Linear Regression	24
2.6.3 Logistic Regression	26
2.6.4 Matrix Decomposition	27
2.6.5 Clustering	28
3 DNA Methylation Heterogeneity Between Phenotypes	30
3.1 <i>RnBeads</i> 2.0: Comprehensive Analysis of DNA Methylation Data	31
3.1.1 Overview of DNA Methylation Analysis Tools	31
3.1.2 Analyzing DNA Methylation Data with <i>RnBeads</i>	32

3.1.3	Application of <i>RnBeads</i> in Cancer and Comparison to Additional Software Packages	39
3.1.4	Discussion	41
3.2	DNA Methylation Dynamics During Aging	42
3.2.1	Estimating DNA Methylation Age in <i>RnBeads</i>	43
3.2.2	DNA Methylation and Aging in Mouse	44
3.2.3	Discussion	45
3.3	Identification of Tissue-Specific and Common Methylation Quantitative Trait Loci in Healthy Individuals Using <i>MAGAR</i>	45
3.3.1	Relationship Between Genotypes and DNA Methylation in MethQTL	45
3.3.2	<i>MAGAR</i> - Methylation-Aware Genotype Association in R	46
3.3.3	Distinct Biological Properties of Tissue-Specific and Common MethQTLs	56
3.3.4	Discussion	62
4	DNA Methylation Heterogeneity Between Samples Sharing a Phenotype	65
4.1	Reference-Free Deconvolution, Visualization, and Interpretation of Complex DNA Methylation Data Using <i>DecompPipeline</i> , <i>MeDeCom</i> , and <i>FactorViz</i>	65
4.1.1	Deconvolution of Complex DNA Methylation Data	65
4.1.2	A Pipeline for Reference-Free Deconvolution of DNA Methylation Data	68
4.1.3	Reference-Free Deconvolution of Lung Adenocarcinoma Data	75
4.1.4	Discussion	81
4.2	Reference-Free Deconvolution of DNA Methylation Data as a Prognostic Tool in Metastatic Melanoma	83
4.2.1	Lack of Predictors for Immune Checkpoint Inhibition Therapy Response in Melanoma	83
4.2.2	Application of the Deconvolution Pipeline on Melanoma Data	83
4.2.3	Prognostic Signature Identified Through Reference-Free Deconvolution	84
4.2.4	Discussion	86
5	DNA Methylation Heterogeneity Within Samples	87
5.1	Quantitative Comparison of Within-Sample Heterogeneity Scores for DNA Methylation Data	87
5.1.1	Within-Sample Heterogeneity in DNA Methylation Data	87
5.1.2	Description of WSH Scores and Data Simulation	89
5.1.3	Application of WSH Scores on Simulated and Experimental Data	97
5.1.4	Discussion	109
6	Conclusions and Outlook	113
6.1	Summary and Perspectives	113
6.2	Outlook	117
	Appendix	119
A.1	Supplementary Figures	119
A.2	Supplementary Tables	125
A.3	Abbreviations	130
A.4	List of Publications	132

A.5 Author Contribution Statements	133
A.6 Copyright Information	133
References	136

List of Figures

1.1	Waddington's epigenetic landscape	2
1.2	The three levels of heterogeneity.	4
2.1	Different levels of chromatin condensation	8
2.2	Open and closed chromatin structures and associated epigenetic marks	9
2.3	The transcriptional initiation process	10
2.4	Methylation of CpG dinucleotides	12
2.5	DNA methylation at gene promoter CGIs	13
2.6	Genome-wide analysis of DNA methylation using WGBS	19
3.1	The <i>RnBeads</i> analysis workflow	33
3.2	DNA methylation heterogeneity in the childhood cancer Ewing sarcoma	40
3.3	Performance of <i>RnBeads</i> and other packages for DNA methylation analysis	41
3.4	Epigenetic age versus chronological age in the UCAM and CEDAR cohorts	44
3.5	Overview of <i>MAGAR</i>	47
3.6	Identifying common and tissue-specific methQTLs	53
3.7	Cell-type specific DNA methylation patterns in the discovery data set	56
3.8	Validating <i>MAGAR</i> and its parameters using simulated data	58
3.9	MethQTL results returned by <i>MAGAR</i>	59
3.10	Common methQTLs identified through colocalization analysis	60
3.11	Validation of tissue-specific and common methQTLs	61
3.12	Properties of shared and tissue-specific methQTLs	62
4.1	Overview of the proposed deconvolution protocol	69
4.2	Overview of covariate adjustment using ICA	72
4.3	Quality control of lung adenocarcinoma data from TCGA	76
4.4	Evaluation of ICA on lung adenocarcinoma data	77
4.5	Selecting the number of LMCs and the regularization parameter for <i>MeDeCom</i>	78
4.6	Interpreting <i>MeDeCom</i> results with <i>FactorViz</i>	79
4.7	Interpreting <i>RefFreeCellMix</i> results with <i>FactorViz</i>	81
4.8	DNA methylation analysis of the melanoma cohort	84
4.9	<i>MeDeCom</i> analysis of the melanoma cohort	85
5.1	Sources of WSH and their manifestation in bisulfite sequencing reads	88
5.2	Simulation setup for modeling WSH	93
5.3	WSH scores in five simulation scenarios	100
5.4	Confusion matrices for the five simulation scenarios	101
5.5	Correlation between simulation parameters and WSH scores	102

5.6	Pairwise comparison of WSH scores on simulated data	103
5.7	Influence of technical biases on WSH scores	105
5.8	WSH scores in the blood cohort dataset	107
5.9	WSH in Ewing sarcoma samples	109
6.1	Integrating genomic, epigenomic, and transcriptomic data	115
A.1	Epigenetic age versus chronological age of young and old mice	119
A.2	Difference between pleiotropy and linkage	119
A.3	Overlapping CpG correlation blocks and tag-CpGs per cell type/tissue.	120
A.4	Description of the validation dataset for methQTL analysis	120
A.5	Quality control for the new samples of the CEDAR cohort	121
A.6	Interpreting <i>MeDeCom</i> results on the Ewing sarcoma RRBS dataset	122
A.7	Elapsed time for the computation of WSH scores	123
A.8	WSH scores in hybrid samples from the DEEP project	124

List of Tables

2.1	Possible outcomes of a binary classification task (confusion matrix).	26
3.1	Parameter settings used for benchmarking tools for DNA methylation analysis. .	36
3.2	Bisulfite sequencing datasets used for the development of a sex classifier.	37
3.3	Overview of <i>MAGAR</i> 's option setting	50
4.1	Overview of published deconvolution tools using DNA methylation data.	67
4.2	Configurable quality filtering options available in <i>DecompPipeline</i>	71
4.3	CpG selection options available in <i>DecompPipeline</i>	73
5.1	Characteristics of different WSH scores	90
5.2	Comparison of average DNA methylation level and WSH scores.	98
5.3	Examples of read configurations and resulting WSH scores.	99
5.4	WSH statistics for different datasets	106
5.5	General guidelines for the application of WSH scores in epigenomic studies. . . .	111
A.1	Feature comparison table of tools for DNA methylation analysis.	126
A.2	Common methQTLs identified using colocalization analysis.	129

Introduction

This thesis comprises six chapters. In this first chapter, I will give an overview of epigenetic regulation and briefly introduce DNA methylation as the main level of epigenetic regulation investigated in the thesis. In the second part of the introductory chapter, I will outline the thesis structure and describe three levels of heterogeneity in DNA methylation data that I will discuss. Additionally, I will formulate research questions answered throughout the work. Since I worked both in a more computer-science-oriented lab and together with wet-lab scientists, I will elaborate on the two aspects of Computational Biology, i.e., developing novel software tools and applying software tools for generating biological hypothesis and insights. The second chapter will introduce the biological and technological background of the thesis and formulate the state of the art. Chapters three to five focus on dissecting the three levels of DNA methylation heterogeneity that we present: between-group, between-sample, and within-sample heterogeneity. In the last chapter, I will summarize the key findings of the thesis, show its implications for the scientific community, and give an outlook on future research directions.

1.1 Epigenetic Regulation

The term *epigenetics*, commonly referred to as inheritable changes of gene expression states that are not encoded by the genome (i.e., the information encoded in the sequence of the deoxyribonucleic acid (DNA)) itself [1], was first conceptualized by Conrad Waddington in 1942 [2]. He introduced the *epigenetic landscape* as a high-dimensional surface defined by epigenetic modifications (Figure 1.1 [3]). The landscape, harbors all potential differentiation states of cells, defines valleys, which represent different cell fates. Cells roll downhill into the valleys of the landscape according to their fate. The lowest points of these valleys represent fully differentiated cell types, while the top of the landscape comprises *pluripotent cells*, i.e., those cells capable of differentiating into all the valleys (cell types) defined. The hills that separate the valleys from one another are defined by epigenetic mechanisms that impede differentiated cells to escape their differentiation valleys. Epigenetic mechanisms define the landscape and are thus crucial drivers of cell-type identity.

Another manifestation of epigenetic regulation manifests in the observation that, although virtually all cells of an organism possess the identical genetic information encoded by DNA, different cell types in the organism differ largely in their functions. For instance, human brain cells use the same genetic information as leukocytes, but possess distinct biological functions. These differences in cellular identity are encoded by the epigenetic program of a particular cell type. Additional manifestations of epigenetic regulation include the inactivation of one of the X-chromosomes in females and *genomic imprinting*, i.e., the parent-of-origin-specific expression of genes. Throughout the last decades and years, the scientific community aimed at illuminat-

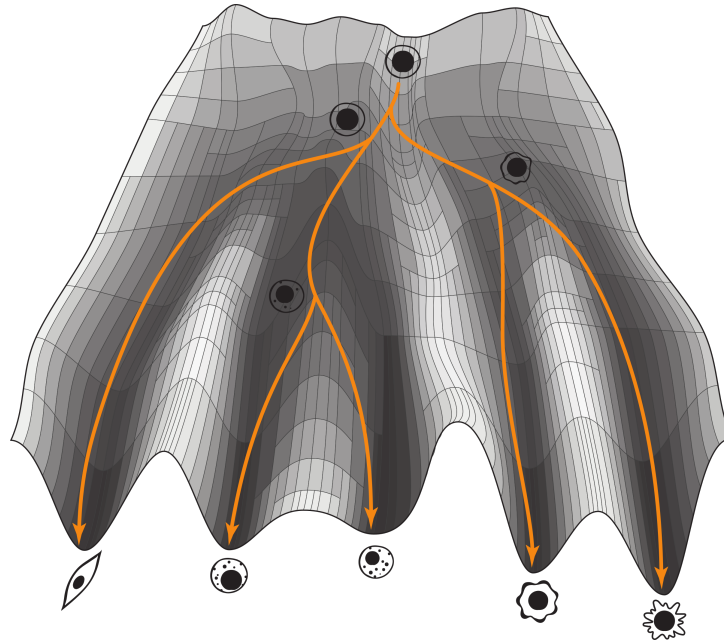


Figure 1.1: Waddington's epigenetic landscape. This high dimensional surface represents different cellular fates according to the orange arrows. Cells traverse various differentiation states, including pluripotent and multipotent progenitors, up to fully differentiated cell types (cells in the bottom of the figure). Modified from <https://doi.org/10.6084/m9.figshare.5285500.v1>, created by Fabian Müller, and inspired from [3]. See Section A.6 for further copyright information.

ing the specific epigenetic patterns that determine cell-type identity, at determining the factors shaping the epigenetic landscape, and at uncovering epigenetic aberrations associated with diseases.

The central dogma of gene expression [4] describes that a protein-coding gene, i.e., a specific segment of the DNA comprising the blueprint of a protein, is first transcribed into messenger ribonucleic acid (mRNA). In a second step, mRNA is translated into a protein that functions as the molecular workhorse of the cell. However, at any point in time, only a subset of the proteins coded for by the genome is needed. To control the expression of genes into proteins, epigenetic mechanisms (including chemical modifications of the DNA and its scaffolding units) regulate the gene expression process. Gene expression is a complex process that involves proteins including transcription factors (TFs), mRNA, and other RNA molecules. The regulation of gene expression occurs both at the transcriptional and at the translational level.

In this work, different levels of epigenetic regulation are addressed. The main focus will be on transcriptional regulation, i.e., on investigating the factors leading to a particular gene being transcribed into mRNA. Notably, protein abundances are also influenced by post-transcriptional and post-translational regulation. Such regulatory mechanisms include alternative splicing and post-translational modifications, which affect tissue-specific protein abundances and functions. Epigenetic regulation at the transcriptional level comprises three major components: DNA methylation, histone modifications, and micro-RNAs (miRNAs), which will be discussed in more detail in Chapter 2.

DNA methylation, the chemical modification of cytosine-guanine dinucleotides (CpGs), is often considered to be the best-studied epigenetic modification. DNA methylation aberrations

have been associated with various diseases, most prominently cancer, where they have emerged as important biomarkers (see Section 2.3). Additionally, DNA methylation is important for establishing cell-type identity and can be measured genome-wide by different technologies (see Section 2.4). In this work, we will particularly focus on DNA methylation, but also investigate different epigenetic layers in healthy and diseased individuals.

1.2 Thesis Outline: Dissecting the Levels of DNA Methylation Heterogeneity

The term *heterogeneity* used in this work refers to differences between the states of a biological system and we will particularly focus on heterogeneity observed in DNA methylation states. After introducing important biological and methodological concepts in Chapter 2, we investigate heterogeneity at three levels (Figure 1.2):

1. Chapter 3 discusses DNA methylation differences between phenotypes
2. Chapter 4 investigates DNA methylation differences between the samples sharing a phenotype (e.g., a group of cancer patients)
3. Chapter 5 dissects DNA methylation differences between different cellular states within a (bulk) sample

Notably, there is no clear-cut distinction between the different levels of heterogeneity, since, for example, high levels of within-sample heterogeneity likely cause elevated between-sample heterogeneity. In contrast to heterogeneity, *homogeneity* of a biological system is rarely investigated, since research is focused on the differences that drive a system to the observed phenotype. We define a *phenotypic group* as a group of individuals that share a phenotype of interest such as a disease or another trait. A sample, however, is obtained from one of the individuals within such a group. We focus on bulk tissue samples, which comprise different cell types and cellular states, and discuss how the emergence of single-cell technology changes how the levels can be dissected in Chapter 6.

Within this work, we refer to the term *Computational Biology* as the development of computational approaches and the application of such approaches to answer biological questions. It is not trivial to translate the results generated by computational tools into biological knowledge. We contributed to that end by using bioinformatic tools developed by ourselves and others on a large variety of biological datasets. In our analyses, we put an emphasis on understanding epigenetic dysregulation associated with diseases.

1.2.1 DNA Methylation Heterogeneity Between Phenotypes

In an epigenome-wide association study (EWAS), different sample groups (e.g., cancer patients and healthy controls) are compared to each other to define CpGs or larger genomic regions (e.g., gene promoters) that are differentially methylated between the groups. This is particularly useful for defining biomarkers of diseases (see Section 2.3), but also for determining the epigenetic changes that contribute to cell differentiation [5]. We need software tools for determining differentially methylated CpGs (DMCs) and differentially methylated regions (DMRs) whose methylation profiles are significantly different between the sample groups defined by

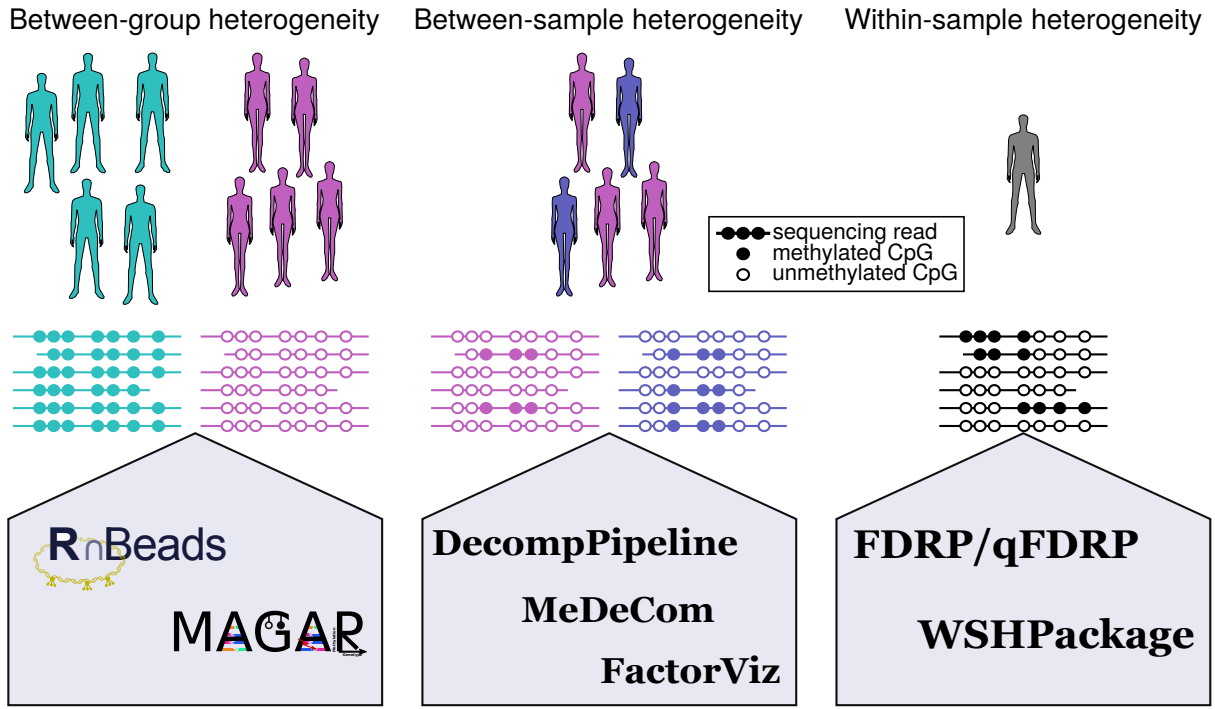


Figure 1.2: The three levels of heterogeneity addressed throughout this work. Between-group heterogeneity shows different DNA methylation states for different phenotypes, between-sample heterogeneity describes DNA methylation differences between samples within a group, and within-sample heterogeneity describes different DNA methylation patterns in an individual sample.

the phenotype. These software tools should fulfill a number of criteria including reliability of the results, transparency of result generation, reproducibility, and reasonable running time. Additionally, the software tools should be easy to use even for non-bioinformaticians to lower the hurdle for in-house data processing. In Chapter 3, we present an extension of such a software package (*RnBeads*) that supports different types of input data and returns a list of DMCs or DMRs. However, *RnBeads* was lacking important features of a DNA methylation analysis, which recently emerged in the scientific community. To that end, we implemented novel methods for detecting differential variability between two groups of samples, for performing genome-wide segmentation, and for imputing missing values. Along with these and other extensions, we conducted a runtime comparison of *RnBeads* to other software tools. The first part of Chapter 3 mainly focuses on presenting novel computational approaches, and showcases the new features of *RnBeads* on a childhood bone cancer cohort.

In contrast, the second part of the chapter is motivated from a more biological perspective. Here, we systematically investigate the influence of organism aging on DNA methylation. We elaborate on epigenetic age prediction in human and mouse, which has been introduced as an additional module in *RnBeads*.

In the third part of Chapter 3, we investigate the relationship between genetic and epigenetic alterations. These genetic alterations (single nucleotide variant or polymorphism (SNV/SNP)) confound the interpretation of identified DMCs, since a genetic alteration can cause an incorrect readout of DNA methylation. We present an in-depth analysis of the relationship between genotype and CpG methylation states by defining methylation quantitative trait loci

(methQTL). So far, there is no software package available implementing a comprehensive workflow for detecting methQTLs from raw genotyping and DNA methylation microarray data. Therefore, we developed a novel computational framework, implemented in the R-package *MAGAR*, to detect methQTLs from genotyping and DNA methylation data. Notably, the package accounts for the properties of DNA methylation data. An open question for the association between genotype and DNA methylation is whether the associations are independent of the cell type or cell-type specific. To answer this question, we analyzed purified blood cells and bowel biopsies. MethQTLs can be used to interpret results of EWAS with respect to genetic modulations, to further understand functional implications of disease-associated genetic alterations, and to illuminate the complex interplay between genotype, DNA methylation, and gene expression.

1.2.2 DNA Methylation Heterogeneity Between Samples Sharing a Phenotype

The second level of heterogeneity investigated in Chapter 4 addresses differences in DNA methylation patterns between samples sharing a phenotype (e.g., control individuals, cancer patients). In addition to heterogeneity introduced by differences in age, sex, or technical biases, cell-type composition substantially contributes to between-sample heterogeneity within a group. DNA methylation patterns are inherently cell-type specific. While this introduces new opportunities for unraveling cell-type identity, it also poses challenges for the analysis of bulk tissue samples and for the interpretation of epigenomic studies. In the context of EWAS, cell-type heterogeneity is typically considered the strongest confounding factor. On the other hand, since cell types can be characterized by their DNA methylation profiles, DNA methylation is a premier candidate for deconvolving complex tissue samples into the constituting cell types. The proportions of the cell types can be associated with disease onset and may be relevant in a clinical setting, where DNA methylation could serve as a disease biomarker or assist in pathological assessment of samples.

Deconvolution of complex DNA methylation datasets is an active research field, since (currently) the majority of epigenomic studies is performed on bulk samples. We define deconvolution as the dissection of a heterogeneous tissue into its main constituents, which can comprise different cell types, but also different sources of variation in the data. However, deconvolution tools, such as *MeDeCom*, require thorough data processing and biological interpretation of deconvolution results remains challenging. To alleviate this problem, we present a three-stage protocol for conducting reference-free deconvolution of complex DNA methylation datasets and an application of the protocol to solid tumors from The Cancer Genome Atlas (TCGA). Similarly, an application of the pipeline on melanoma samples of patients treated with immune checkpoint inhibition (ICI) therapy substantiates the clinical use of deconvolution analysis.

1.2.3 DNA Methylation Heterogeneity Within Samples

While the first two levels of heterogeneity in epigenomic data are commonly addressed for, the third one – heterogeneity within an individual sample – is often neglected. More specifically, DNA methylation of a single DNA strand is truly binary, i.e., the cytosine is either methylated or unmethylated. Thus, without any form of heterogeneity within a sample, all CpG dinucleotides measured in an epigenomic study would be binary. However, some CpGs measured either show some minor deviations from this binary state (e.g., 96% methylation), which can

be mainly attributed to technical artifacts. In addition, biological variations (e.g., cell-type heterogeneity, allele-specific methylation) contribute to within-sample DNA methylation heterogeneity (WSH).

Information about DNA methylation states of individual molecules is preserved in the raw sequencing reads (see also Figure 1.2). In Chapter 5, we systematically address different sources of within-sample heterogeneity and discuss how this heterogeneity can be quantified using genome-wide within-sample heterogeneity (WSH) scores. A comprehensive benchmark of different WSH scores is currently missing. Thus, we present a systematic benchmark of existing metrics and introduce a novel score for quantifying within-sample heterogeneity from bisulfite sequencing data, *qFDRP*. WSH scores, such as *qFDRP*, can be used for interpreting DMRs, for segmenting the genome into regions with high and low DNA methylation heterogeneity, and for predicting tumor purity.

Background

In this chapter, I will introduce epigenetic regulation in more detail, explain important terms used throughout the thesis, and report the state of the art. I will introduce the key players of epigenetic regulation, with a focus on DNA methylation as one of the most extensively-studied epigenetic marks. I will also discuss associations between DNA methylation aberrations and diseases, and I will point out potential implications for the clinical use of epigenetic markers. Furthermore, I will present technologies for mapping DNA methylation genome-wide and outline the routine, low-level processing of the data. This chapter requires a basic understanding of the terms used in molecular biology, and will not give a comprehensive list of definitions for all the terms used. In the last part of the chapter, I will introduce key concepts of statistical learning and computational methods that will be used throughout this work.

2.1 Mechanisms of Epigenetic Gene Regulation

As discussed in Chapter 1, epigenetic mechanisms are crucial drivers of cellular identity and important for X-chromosomal inactivation and genomic imprinting. Determining which genes, i.e., stretches of DNA coding for a protein or RNA, are transcribed and translated at a particular point in time is critical for organism development. Transcriptional regulation is a complex process that influences which sequence of the DNA is transcribed into mRNA by the protein RNA Polymerase II (RNAPII).

2.1.1 Chromatin Structure

The human DNA comprises more than 3 billion base pairs (bp), which sum up to a total length of 2 m of linear DNA per diploid cell. Thus, the DNA molecule needs to be highly condensed to fit into the cell nucleus, which is typically 5 μm in diameter for most animal cells [6] and about 10 μm for most human cells. To accomplish this high level of compression, DNA is wrapped around *nucleosomes* – protein complexes comprising eight histone subunits [7]. 146-147 bp of genomic DNA are wrapped around the histone octamer, and nucleosomes are further compacted into higher-order structures to form chromosomes (Figure 2.1). The entirety of the chromosomes, the chromatin, exists in two functional states that are characterized by different levels of DNA condensation. In the most dense state, nucleosomes are tightly packed and a large fraction of the DNA is wrapped around histone octamers. This *heterochromatic* state reaches the highest level of compression during the metaphase of the cell cycle, in which chromosomes are microscopically detectable (Figure 2.2). Heterochromatic structures compromise protein binding to DNA in general and specifically impair the binding of RNAPII and transcription factors (TFs). The second state of chromatin, *euchromatin*, is characterized by a loosened structure, in

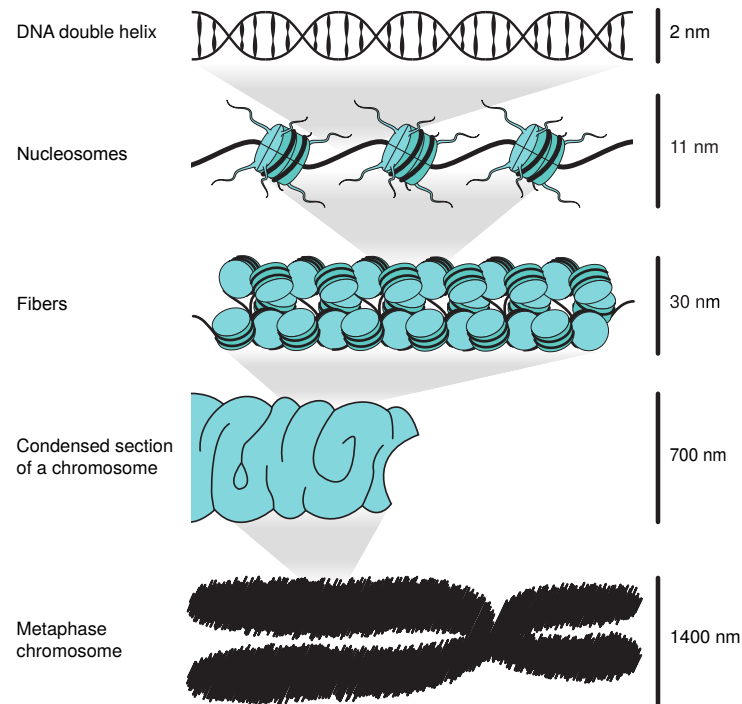


Figure 2.1: Different levels of chromatin condensation. DNA is wrapped around histone octamers to form nucleosomes. These are further condensed into fibers to form chromosomes during the metaphase of the cell cycle. Modified from <https://doi.org/10.6084/m9.figshare.5285488.v1>, created by Fabian Müller.

which larger fractions of DNA are not wrapped around nucleosomes. Thus, DNA in euchromatin is more accessible to proteins than DNA in heterochromatin and highly expressed genes are located within euchromatic regions of the genome.

2.1.2 Regulatory Elements of Gene Expression

The genome comprises different regulatory elements, which determine the set of genes expressed. To initiate transcription of a particular gene into mRNA, the transcription machinery – a protein complex – is recruited around the transcription start site (TSS) in the gene promoter region (Figure 2.3). The binding of RNAPII to the promoter region of a gene is facilitated by the recruitment of proteins (TFs). Binding of the transcriptional initiation complex (i.e., RNAPII and TFs) to the promoter region requires accessible DNA, i.e., euchromatin (Figure 2.2). In addition to proximal elements regulating gene expression, such as promoters, more distal regulatory elements, including enhancers, impact transcriptional initiation (Figure 2.3). Additional TFs can be recruited to enhancer elements, which form complexes with the proteins binding to the promoter region, thus allowing for the initiation of transcription. Enhancer elements are classified according to their genomic distance to the TSS as proximal or distal enhancers, respectively. Proximal enhancers are regions directly flanking the promoter region according to the *Ensembl Regulatory Build* [8], while distal enhancers are regions further away from the TSS. Active enhancers are marked by histone modifications of the N-terminal tails of the histones such as H3K27ac and H3K4me1 (further explained in Section 2.1.3) and characterized by a low DNA methylation level of the neighboring CpGs. Enhancers harbor transcription factor bind-

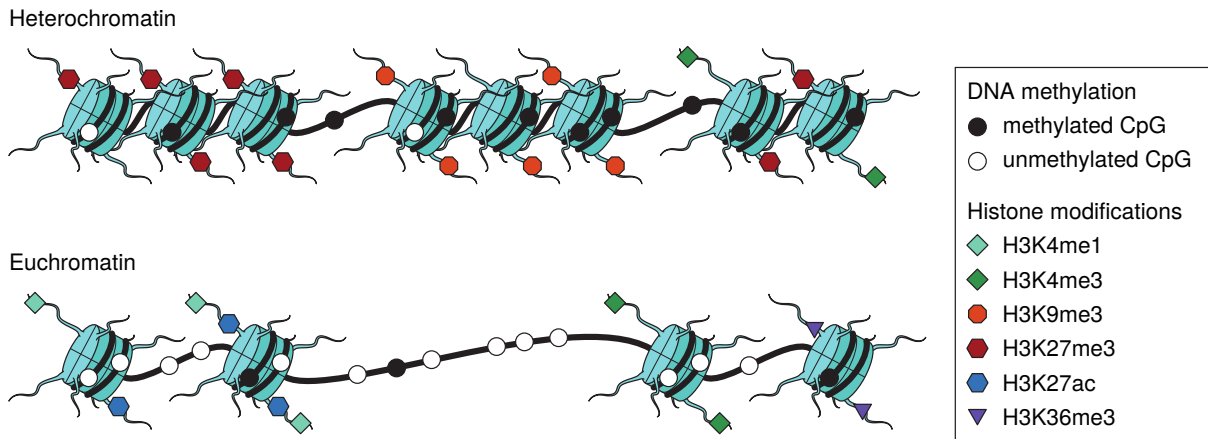


Figure 2.2: Open and closed chromatin structures and associated epigenetic marks. H3K9me3 and H3K27me3 are histone marks that are associated with closed chromatin (heterochromatin) together with high levels of CpG methylation. H3K4me1, H3K27ac, H3K36me3, and low levels of CpG methylation are associated with open chromatin regions (euchromatin). Obtained from <https://doi.org/10.6084/m9.figshare.5057566.v1>, created by Fabian Müller.

ing sites (TFBS) for TFs that form a complex with RNAPII. Notably, transcriptional initiation is a complex process, which can be modulated by further epigenetic and other factors. The target genes of an enhancer often remain elusive and determining them is subject to active research [9]. Given the local epigenetic pattern, gene expression levels can be reliably predicted [10] and changes in the epigenome can alter gene expression levels in a disease setting. In addition to enhancer elements, genomic regions called insulators influence chromatin contacts of distant (in sequence space) genomic regions. Insulators typically carry binding sites for the transcription factor CCCTC-binding factor (CTCF) and form the boundary between euchromatin and heterochromatin [11]. The binding of CTCF influences the contact between an enhancer element and its target gene.

Gene expression levels, i.e., the number of mRNA molecules transcribed from a gene, are often referred to as the functional readout of epigenetic patterns in a particular environment and are an indication of the functional effect of transcriptional regulation. Throughout this work, gene expression levels, typically measured by the abundances of mRNAs at a particular point in time via RNA sequencing (RNA-seq), will be used. Epigenetic patterns regulate gene expression at the transcriptional level and thus modulate mRNA abundances. These mRNAs are translated into proteins to perform cellular functions, and the abundances of mRNA molecules are important indicators for the activity levels of the proteins in a cell. Post-transcriptional regulation and modifications will not be discussed in this thesis, but rather the transcriptional regulation through epigenetic marks, such as histone modifications and DNA methylation, will be. Additionally, gene expression regulation by small RNA molecules, including miRNAs and small interfering RNAs (siRNAs) is considered the third basic epigenetic mechanism, but will not be addressed in this work.

2.1.3 Histone Modifications

Epigenetic modifications co-occur with open (euchromatin) and closed (heterochromatin) chromatin and determine chromatin structure. DNA is accessible for TFs and RNAPII in euchro-

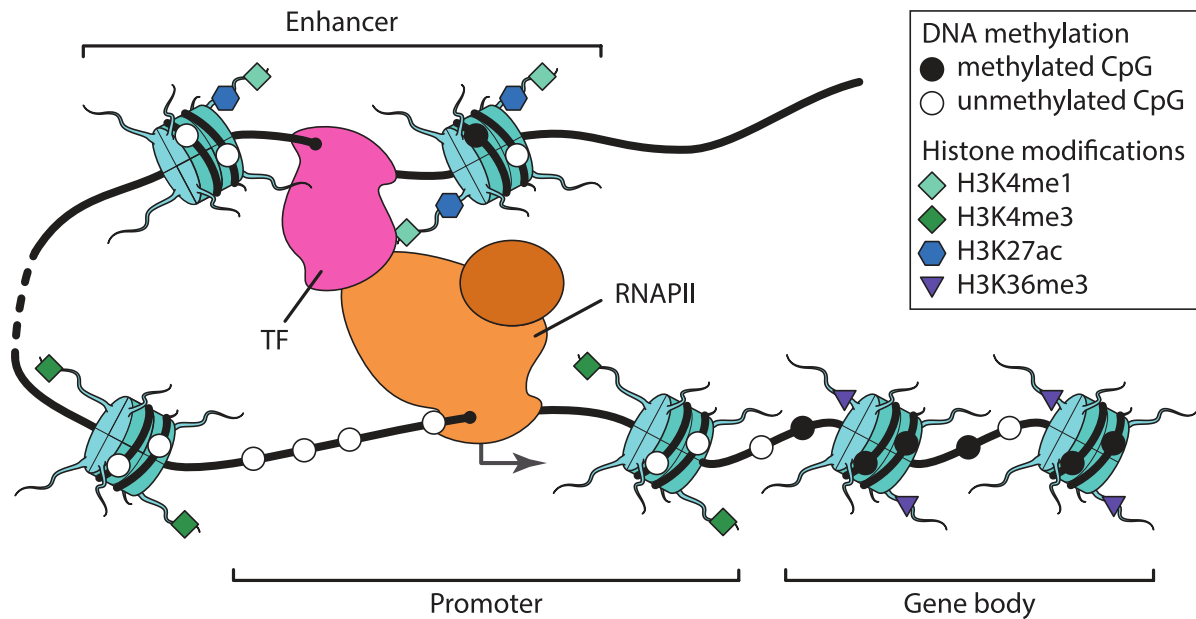


Figure 2.3: The transcriptional initiation process. RNA Polymerase II (RNAPII) binds an unmethylated promoter region along with TFs recruited to form the transcriptional initiation complex. These TFs can bind at distal or proximal genomic elements (enhancers) and form a chromatin loop toward the promoter complex. Both the enhancer and the promoter region generally require an open chromatin structure, which is regulated by epigenetic mechanisms, such as DNA methylation and histone modifications. Modified from <https://doi.org/10.6084/m9.figshare.5285473.v1>, created by Fabian Müller.

matin and inaccessible in heterochromatin. Thus, it is crucial to determine which epigenetic marks are associated with open and closed chromatin. The N-terminal tail of the histone protein H3 (a subunit of the histone octamer) is subject to chemical modifications, which modulate chromatin openness and closeness [12]. These modifications are referred to as *histone modifications*, *histone marks*, or *chromatin marks*. For instance, H3K27me3 refers to an addition of three methyl groups to the 27th lysine (counted from the N-terminus) of histone H3 (Brno nomenclature [13]). H3K27me3 and H3K9me3 are associated with closed chromatin, while H3K4me1 and H3K9ac (lysine acetylation) are associated with euchromatin (Figure 2.2, [14]). The chemical modification impacts the binding affinity of DNA to the histones and influences the compression level of chromatin. Many histone modifications have been identified and their associations with the chromatin structure and gene expression state have been investigated [15]. Within this work, histone modifications will be used to segment the genome into different functional units.

Epigenetic regulation is often more complex than the simplistic correlation between a single chromatin mark and the chromatin state. Rather, it is an interplay between different chromatin marks, DNA methylation, the local sequence context, and TF binding that modulates local chromatin accessibility. One example is presented by *bivalent* domains, which harbor both repressing (e.g., H3K27me3) and activating (H3K4me3) histone marks. Computational methods such as *ChromHMM* [16] use histone modification data generated using Chromatin Immunoprecipitation sequencing (ChIP-seq [17]) to segment the genome into different chromatin states. Such segmentation methods exploit the combined information of multiple histone marks. The segmentation can be substantially improved using further epigenetic information such as DNA

methylation data [18].

Determining which regions of the chromatin are accessible to proteins is crucial for identifying potential regulatory regions and actively expressed gene. Thus, epigenomic techniques for measuring DNA accessibility have been developed for detecting euchromatic and heterochromatic structures. Those methods are based on enzymes that particularly target regions of open chromatin. Most notably, DNaseI-sequencing (DNaseI-seq) [19], ATAC-sequencing (ATAC-seq) [20], and NOMe-sequencing (NOMe-seq) [21] can be used to map regions of accessible chromatin. In addition, with the Hi-C technique [22] one can measure interactions between different chromosomal regions to define *topologically associated domains* (TADs)¹ and euchromatic as well as heterochromatic regions at relatively low resolution. In order to illuminate chromatin contacts for specific genomic regions such as gene promoters, promoter-capture Hi-C can be used to map the contacts at a higher resolution, which can also be used to connect putative regulatory elements (e.g., enhancers) to their target genes [9].

2.1.4 DNA Methylation

DNA methylation is a widely-studied epigenetic mark. In mammalian genomes, cytosine bases in the context of a CpG dinucleotide, i.e., a cytosine followed by a guanine in the DNA sequence, can be chemically modified by the addition of a methyl group to the fifth carbon of the cytosine ([24], Figure 2.4). The resulting base is referred to as 5-methylcytosine (5mC) and has distinct chemical and biological properties compared to an unmodified cytosine. Notably, CpG dinucleotides are symmetric (they occur on both the forward and reverse DNA strand) and the methyl group is predominantly present on both DNA strands. Evolutionary this led to the biased distribution of DNA methylation in CpG dinucleotides, since methylated cytosines mutate to thymines at a high rate [25, 26]. In addition to CpG methylation, other bases such as adenine can be methylated, which is frequently found in bacteria [27] and plants [28]. In this work, CpG methylation will be discussed in human and mouse.

CpG dinucleotides are heavily depleted in the human genome [29] due to their high mutation rate. A notable exception is presented by short regions of around 1,000 bp that exhibit a significantly elevated CpG density. These regions are called CpG islands (CGIs), which occur in around 70% of all human gene promoters [30]. While CpGs outside of CGIs are generally methylated, CpGs within CGIs are unmethylated. The methylation states of neighboring CpGs (distance below 1 kilobase (kb)) are typically highly correlated [31], since DNA methylation is likely regulated by regional rather than local mechanisms.

Establishment, Maintenance, and Removal

Methylation of CpGs in the human genome is mainly regulated by two classes of enzymes: DNA methyltransferases (DNMTs) add a methyl group to an unmethylated cytosine, while ten-eleven translocation (TET) enzymes facilitate the removal of the methyl group. DNMTs use S-adenosylmethionine (SAM, [32]) as the donor of methyl groups and are subdivided according to their function into the maintenance (DNMT1) and *de-novo* class (DNMT3A, DNMT3B, DNMT3L). DNMT1 preferentially modifies asymmetrically (only on one strand) methylated CpG dinucleotides and copies the DNA methylation state of the methylated cytosine to the

¹A TAD is a genomic region, in which the frequency of physical contacts between any pair of stretches of DNA is higher than the frequency of interactions to stretches of DNA outside of the TAD [23].

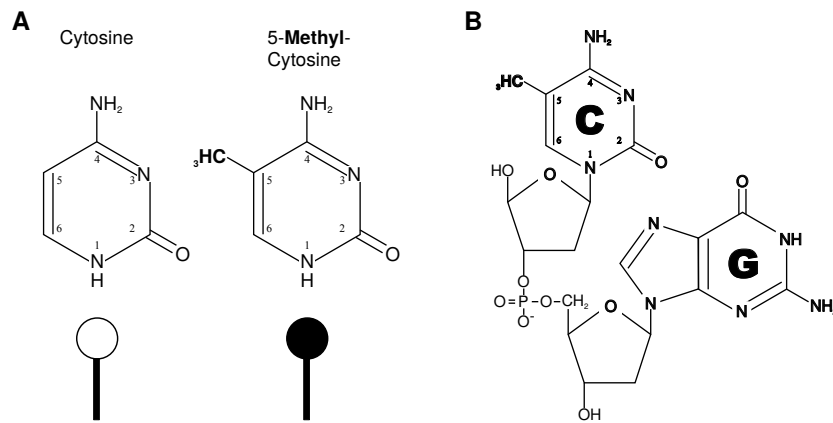


Figure 2.4: Methylation of CpG dinucleotides. **A:** Cytosine bases can be methylated by the addition a methyl group to the fifth carbon position. Filled circles represent methylated, while unfilled circles represent unmethylated cytosines. **B:** DNA methylation almost exclusively occurs in CpG dinucleotides.

cytosine on the complementary DNA strand. Thus, DNMT1 is essential for copying the methylation signature to the newly synthesized DNA strand after replication and is crucial for the survival of human embryonic stem cells (ESCs, [33]). DNMT3A and DNMT3B preferentially target unmethylated CpG dinucleotides and introduce methylation *de-novo*. The process is catalyzed by DNMT3L, which does not have a methyltransferase activity on its own [34].

DNA methylation is either passively lost during replication, when DNMT1 fails to copy the methyl group to the newly synthesized DNA strand (DNA methylation erosion) or actively removed. The active removal is catalyzed by TET enzymes, which oxidize 5mC to 5-hydroxymethylcytosine (5hmC). Subsequent oxidation products include 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). The latter two cytosine variants are actively removed from the DNA strand using either the base excision repair (BER) machinery or by thymine DNA glycosylase (TDG) and are replaced by an unmodified cytosine. The oxidation products of 5mC (5hmC, 5fC, 5caC) are substantially less abundant than 5mC in most mammalian genomes, whereas 5hmC levels can reach up to 40% of the 5mC levels in mouse brains [35]. The functional roles of the modifications of 5mC are only beginning to be understood [36]. For instance, high levels of 5fC have been reported at enhancer elements in mouse [37], and 5hmC has distinct functions in comparison to 5mC at human gene promoters [38].

Function

DNA methylation has been implicated in functions within several biological processes. Generally, high levels of DNA methylation at CpG dinucleotides are considered a repressive mark for gene expression. Methylated CpGs are associated with heterochromatin formation through the recruitment of methyl-CpG-binding domain (MBD) proteins, which themselves recruit chromatin remodelers. This function is especially relevant for transposable elements, which generally show a high level of DNA methylation. Thus, DNA methylation represses the expression of transposable elements and contributes substantially to genome stability [39]. Furthermore, DNA methylation is essential for genomic imprinting [40] and X-chromosomal inactivation [29].

In contrast to CpGs located in genomic regions of low CpG density, CGIs are preferentially

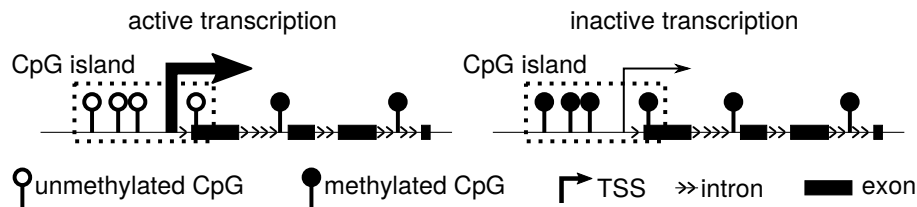


Figure 2.5: DNA methylation at gene promoter CGIs. Low levels of DNA methylation at CGI gene promoters are an indication of active transcription, while higher levels of DNA methylation of the CGI repress transcription.

unmethylated. The methylation states of CGIs located at promoter regions are indicators of the transcriptional state of the gene. High levels of DNA methylation are associated with inactive transcription and low levels of DNA methylation are an indication of an active transcriptional state (Figure 2.5). In addition to this negative association between DNA methylation and gene expression, high levels of CpG methylation throughout the gene body of a transcribed gene are an indication of active transcript elongation [41]. An important characteristic of DNA methylation is its cell-type specificity. DNA methylation patterns can be used to discern between different cell types and cellular subtypes [5]. Due to its quantitative readout as the overall methylation level of a particular CpG in a bulk tissue sample, DNA methylation is a premier candidate for epigenomic deconvolution of complex tissues (discussed in Chapter 4).

Development and Aging

DNA methylation is particularly dynamic in the early stages of life. More specifically, genomes undergo two phases of global demethylation, followed by establishment of the DNA methylation states [42]. First, DNA methylation is fully erased during gamete formation, followed by the establishment of DNA methylation in a parent-specific manner (genomic imprinting, [40]). Second, DNA methylation in the zygote is removed first from the paternal and then from the maternal copy of the genome [43]. This second round of demethylation is followed by setting the DNA methylation pattern of the developing organism. After birth, the global DNA methylation pattern undergoes larger changes up to adolescence. In the later years of life, the methylome, i.e., the genome-wide profile of DNA methylation, is less affected by changes, such that DNA methylation patterns of 80-year-old people are nearly indistinguishable from those of centenarians [44, 45].

Since virtually all organisms are affected by aging, this process is a prominent research target for epigenetic research. Particularly, understanding the human aging process is relevant for developing anti-aging interventions and for prolonging the period of a healthy human life. DNA methylation has recently emerged as a powerful biomarker for the human aging process. Generally, the genome-wide DNA methylation level decreases with age, while the characteristic, strong correlation between adjacent CpGs is reduced in centenarians [45]. Furthermore, DNA methylation levels at particular CpGs can be used as reliable predictors of the chronological age of healthy individuals (see Section 3.2, [46, 44, 47]).

2.2 Influence of Genetic Variation on Epigenetic Regulation

After the publication of the sequence of the human genome in the year 2001 [48, 49], research focused on unraveling genetic variations between individuals associated with diseases or other complex traits. Single nucleotide polymorphisms (SNPs) are sequence variations from the reference genome at individual genomic positions that occur in more than 1% of the human population [49]. To determine disease-related genetic variations, genome-wide association studies (GWAS) associate SNPs to a phenotype of interest (e.g., a disease) by comparing a discovery cohort, i.e., a group of individuals sharing the trait of interest, to a group of control individuals. Using this methodology, SNPs linked to, e.g., schizophrenia, type 2 diabetes, and rheumatoid arthritis [50, 51, 52] have been identified.

Genetic alterations have also been associated with quantitative traits in order to define quantitative trait loci (QTL). Such QTLs have been identified, among others, for gene expression levels (expression QTL, eQTL [53, 54]), chromatin accessibility (caQTL [55]), metabolomic alterations (metabolomic QTL, mQTL [56]), and DNA methylation (methylation QTL, methQTL, [57, 58]). A QTL is defined as a SNP whose occurrence significantly correlates with the quantitative trait investigated in the population of study participants. For instance, a methQTL is a genetic variant that correlates with the DNA methylation state of an individual CpG or of multiple, jointly-regulated CpGs. The combination of multiple QTLs for different traits allows for the characterization of the interplay between complex traits, epigenetics, and genetics. For instance, a CpG methylation state can influence the expression level of a gene or *vice versa* and the genetic associations serve as a mediator for inferring causality from the interactions [59]. Finally, QTLs can be compared with SNPs associated with a disease (often referred to as a GWAS hit) to determine the functional role of the quantitative trait in the disease [60, 61]. MethQTLs will be discussed in more detail in Section 3.3.

2.3 DNA Methylation Aberrations and Their Association With Human Diseases

DNA methylation has emerged as an important biomarker for various diseases, since it is comparatively easy to measure and more stable with regard to environmental influences than, for instance, gene expression. Due to the reliability and sensitivity of the Illumina BeadArrays and more local approaches such as pyrosequencing (see Section 2.4), DNA methylation is becoming increasingly relevant for clinical diagnostics and will contribute to precision medicine. DNA methylation aberrations can cause transcriptional dysregulation, which can have severe systemic effects for the affected organism. Epigenetic therapies can include modulation of the DNA methylation patterns by targeting *DNMT* and *TET* genes using genetic modification tools, such as CRISPR/Cas9 or targeted demethylation using 5-azacytidine [62].

2.3.1 DNA Methylation Aberrations in Human Diseases

In addition to its connection to healthy human aging [44, 45], cigarette smoking [63, 64, 65], and obesity [66], DNA methylation has been implicated in multiple diseases through epigenome-wide association studies (EWAS). While DNA methylation states of a subset of CpGs change gradually with healthy human aging, a connection between aberrant DNA methylation pat-

terns and diseases causing premature aging, such as Hutchinson-Gilford Progeria and Werner syndrome, has been detected [67]. Further EWAS revealed associations of DNA methylation with rheumatoid arthritis [68], schizophrenia [69], and inflammatory bowel disease [70]. Similarly, DNA methylation is a mediator of genetic risk for rheumatoid arthritis [68] and is associated with Crohn's disease and ulcerative colitis [57]. Moreover, characteristic differences in DNA methylation patterns have been detected for brain samples from multiple sclerosis patients and healthy controls [71], as well as in twins clinically discordant for multiple sclerosis [72]. Additional associations between DNA methylation and diseases have been reported for major depression [73], and type 1 diabetes [74]. However, the detected associations are correlative, and a causal connection between DNA methylation aberrations and disease onset remains to be investigated for most diseases [75]. Thus, DNA methylation is currently almost exclusively employed as a biomarker instead of as a potential therapy target.

2.3.2 DNA Methylation Aberrations and Cancer Progression

DNA methylation aberrations have been intensively investigated in cancer progression, since cancers exhibit largely disordered DNA methylation patterns throughout the genome. Most notably, CGIs become hypermethylated, while otherwise highly methylated regions in the genome become hypomethylated. This phenomenon is referred to as the *CGI Methylator Phenotype* (CIMP) [76]. Hypermethylation of CGIs within tumor suppressor genes, such as *p53*, can promote cancer progression. On the other hand, transposable elements become hypomethylated, which hampers the repressive function of DNA methylation on transposable element expression. This process contributes to increased genome instability and chromosomal rearrangements. The *p53* gene acts as an important regulator of the DNA methylome, and its altered expression is linked to both loss and gain of DNA methylation [77]. While loss of DNA methylation is distributed across different regions of the genome, hypermethylation occurs more focally [78]. In depth functional characterizations of DNA methylation alterations have been performed for prostate cancer, hepatocellular carcinoma, acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), and glioblastoma [79, 80, 81, 82, 83]. DNA methylation can be used as an effective biomarker for diagnosis and prognosis [84], but also substantially facilitates pathological classification of brain tumors [85]. The latter classification has entered routine clinical practice. Due to specific DNA methylation aberrations in cancers, DNA methylation remodelers, including DNMTs, are a premier target for drug development and have been successfully applied for the treatment of solid tumors [86].

2.4 Genome-Wide Mapping of DNA Methylation

DNA methylation is a well-studied epigenetic mark, since it can be quantified genome-wide using various experimental assays. Notably, it is important to determine the fraction of molecules representing methylated CpGs over all molecules assayed, i.e., computing the DNA methylation level or beta value. Two types of technologies for assaying DNA methylation are commonly used in the scientific community. BeadChip Arrays are microarrays that measure a subset of the CpGs in the human genome using primer extension and have color intensities as the readout. Second, sequencing-based approaches use high-throughput sequencing to quantify the number of methylated molecules at a genomic location. Both methods inherently rely on bisulfite

conversion of unmethylated cytosines to uracils. Sodium bisulfite converts unmethylated cytosines into uracils through sulfonation and subsequent deamination and desulfonation, while methylated cytosines are protected from this reaction and remain as cytosines after the conversion [87, 88]. Using this technique, the chemical modification present on the DNA is converted into a sequence alteration that can be detected using standard genomic sequencing or microarray technology. The resulting data can be analyzed using computational pipelines (see Section 2.5). While ATAC-seq and ChIP-seq data can be used to determine regions in the genome being in a particular chromatin state using peak-calling algorithms [89], the DNA methylation assays discussed here provide a quantitative methylation level for each CpG.

2.4.1 Illumina Infinium BeadChip Arrays

Since the human genome comprises around 28 million CpG dinucleotides, high sample numbers are required for detecting reliable differences between phenotypes. To facilitate large-scale analysis of DNA methylation, Illumina proposed the Infinium[®] Methylation BeadArray series. BeadArrays are microarrays that comprise several thousands of probes harboring CpG dinucleotides to be routinely assayed. The first version of the Infinium microarrays – the Illumina Infinium HumanMethylation27 (27k) bead array – assays more than 27,000 CpGs. The newer generations of the series – the Illumina Infinium HumanMethylation450 (450k) and the Illumina Infinium EPIC (EPIC) arrays – comprise more than 450,000 and more than 850,000 CpGs, respectively. While the 27k and the 450k arrays mostly comprise CpG sites in gene promoters of cancer-related genes, the newer generation also focuses on CpGs in putative enhancer regions.

For each of the CpG sites of interest, the bead array comprises multiple beads with multiple probes harboring a sequence flanking the CpG site. The bead arrays rely on primer extension of the probe sequence using the bisulfite converted DNA as a template. The deoxyribose nucleoside triphosphates (dNTPs) used as substrate to the primer extension carry either a green or red fluorescence label. There are two types of beads and corresponding probes available on the newer generations (450k and EPIC) of the chip: type I probes employ two probes per locus – an unmethylated and a methylated one [90]. Since the binding of an unmethylated target sequence (i.e., a thymine after bisulfite conversion) to a methylated probe leads to the termination of the elongation of the probe sequence (and *vice versa* for methylated target sequence and unmethylated probe), type I probes can be measured using a single color channel. In contrast, type II probes comprise a generic bead type that binds to both methylated and unmethylated sequences. Thus, type II probes require two color channels for the different binding events corresponding to a methylated/unmethylated target sequence. Finally, both probe types emit light signals for the methylated and the unmethylated channels, respectively [90, 91], which can be scanned using a Illumina HiScan or iScan machine. According to Illumina's recommendation, 250 ng of genomic DNA is required as input to the EPIC array². From these light intensities, methylation values are called as beta values through:

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + \alpha} \quad (2.1)$$

where M is the signal intensity of the methylated channel and U the signal intensity of the unmethylated channel. The range of values for M and U is in the 1,000s to 10,000s and the parameter α is a constant offset that is typically set to 100 [90] for addressing the issue of both M

²<https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html>

and U being very small. In addition to CpGs, quality control probes are available for checking for the technical quality, including hybridization, specificity, and bisulfite conversion. Additionally, a few dozen genomic loci harboring SNPs with high minor allele frequency (MAF) have been established on the microarray, which allow for the identification of potential sample mix-ups in a genetically matched design.

Due to the difference in design, type I probes tend to yield a larger dynamic range of beta-values [92]. Thus, normalization methods are required to match the distributions of the two types of probes and to avoid spurious associations with a phenotype. Multiple methods are available such as quantile normalization [93] or functional normalization [94] (more details in Section 2.5). The Illumina BeadArrays can reliably detect small methylation differences and is in clinical use [85]. Since the bead arrays reliably and reproducibly return methylation values for the CpGs available, they allow for combined analysis of datasets generated at different places.

Using the same technology and replacing the methylation-aware probes by different genetic variants, Illumina provides microarrays, such as the Illumina Infinium OmniExpress or the Illumina Infinium OmniExpressExome BeadArray, for genotyping around 750,000 and one million SNPs, respectively. These microarrays have been used for GWAS and for associating genetic variants with complex traits (see Section 3.3).

2.4.2 Sequencing-Based Approaches

Due to the decreasing cost of high-throughput sequencing, bisulfite sequencing has become another frequently applied method for profiling DNA methylation. Next-generation sequencing (NGS) uses the *sequencing-by-synthesis* strategy established by Illumina to yield *sequencing reads*. The process is conducted in a sequencing machine, such as the Illumina HiSeq2500 or the Illumina NovaSeq. The fragmented input sample (sequencing library) is hybridized to molecules on a sequencing *flow cell* using a sequencing adapter, and clusters of fragments are generated through amplification. Multiple samples can be analyzed on the same flow cell by using a unique barcode (index) per sample. During the sequencing process the integration of color-labeled dNTPs into the template sequence is captured using an ultra-high resolution camera. The resulting sequencing read is a sequence of letters that represents the nucleotides called (A, C, T, or G). Newer sequencing technologies (third-generation sequencing), including Oxford Nanopore sequencing [95] and Pacific Biosciences' (PacBio) single molecule real-time (SMRT, [96]) sequencing, allow for generating longer sequencing reads of 10s to 100s of kilobases in comparison to the typical 50-250 bp read length generated by Illumina sequencing. Importantly, third-generation technologies tend to have higher error rates than Illumina sequencing and are more expensive. Fragments can either be sequenced from one side only (single-read sequencing) or from both directions together (paired-end sequencing).

In bisulfite sequencing, unmethylated cytosines are converted to uracils, which themselves are replaced by thymines in the polymerase chain reaction (PCR). In the resulting sequencing read, a methylated cytosine is represented as a cytosine, while the unmethylated cytosines appear as thymines. Analogously to the beta-value defined for bead arrays, the methylation level of a CpG is computed as:

$$\text{methylation level}_{\text{CpG}} = \frac{\#C}{\#C + \#T} \quad (2.2)$$

where $\#C$ is the number of sequencing reads supporting a (methylated) cytosine and $\#T$ the number of reads supporting a thymine (i.e., an unmethylated cytosine). A parameter similar to

α in Equation 2.1 is not required, since positions with both $\#T$ and $\#C$ being zero are not considered. There are two frequently used methods for bisulfite sequencing: Whole-Genome Bisulfite Sequencing (WGBS) and Reduced-Representation Bisulfite Sequencing (RRBS). Enrichment-based approaches are less frequently employed for quantifying DNA methylation genome-wide. Notably, in contrast to the Illumina BeadArrays, bisulfite sequencing can be performed for any species provided that a reference-genome is available. Thus, this technology is the current standard for investigating DNA methylation in mice, although Illumina recently introduced a microarray for murine samples³.

In contrast to these genome-wide approaches, several methods for mapping DNA methylation at specific genomic regions have been introduced. For instance, mass spectrometry can differentiate between methylated and unmethylated molecules and does not rely on bisulfite conversion, but cannot be used to measure DNA methylation genome-wide. Similarly, local deep sequencing using, e.g., the Illumina MiSeq technology can be used to comprehensively characterize short genomic regions of size less than 500 bp with sequencing depths up to several thousand reads [97]. Alternatively, pyrosequencing or methylation-specific PCR can be used to assay DNA methylation of short genomic regions. However, genome-wide approaches applied to human samples will be in the focus for this thesis.

Whole-Genome Bisulfite Sequencing

Whole-Genome Bisulfite Sequencing (WGBS, [98]) is typically considered the gold-standard method for genome-wide analysis of DNA methylation. WGBS assays virtually all CpGs in the human genome (typically more than 90%) at a lower read depth than local deep sequencing, yet providing reliable estimates of CpG methylation states.

The input to WGBS is cell material that can be obtained from any tissue, purified cell types, or from cell cultures. Typically, many cells are required to yield a sufficient amount of DNA and cell populations (bulk tissue samples) are frequently assayed. As a first step of the WGBS protocol, genomic DNA is extracted from the cells (Figure 2.6). In the most sensitive protocol proposed to date, tWGBS [99], only 1 ng of genomic DNA is required to yield a sequencing library. After fragmentation of the genomic DNA, the fragments are ligated to sequencing adapters and subsequently treated with sodium bisulfite. The obtained fragments are sequenced using a NGS machine. The final steps of the protocol involve bioinformatic processing of the data, which will be discussed in more details in Section 2.5.2. While single-cell bisulfite sequencing methods become increasingly available [100, 101, 102], they still suffer from high sequencing costs and low genomic coverage [103], and they generate sparse data matrices.

Reduced-Representation Bisulfite Sequencing

A major disadvantage of WGBS is that many fragments are sequenced, which comprise only few or no CpG dinucleotides and thus do not yield methylation information. This is due to the uneven distribution and due to the overall depletion of CpGs in the human genome. Additionally, many CpGs in the human genome are constantly methylated across virtually all cellular states and cell types. Thus, Reduced-Representation Bisulfite Sequencing (RRBS) has been developed as a cost-effective alternative for obtaining DNA methylation information for 10-20% of all CpG dinucleotides in the human genome [104]. The RRBS protocol employs enrichment

³<https://emea.illumina.com/products/by-type/microarray-kits/infinium-mouse-methylation.html>

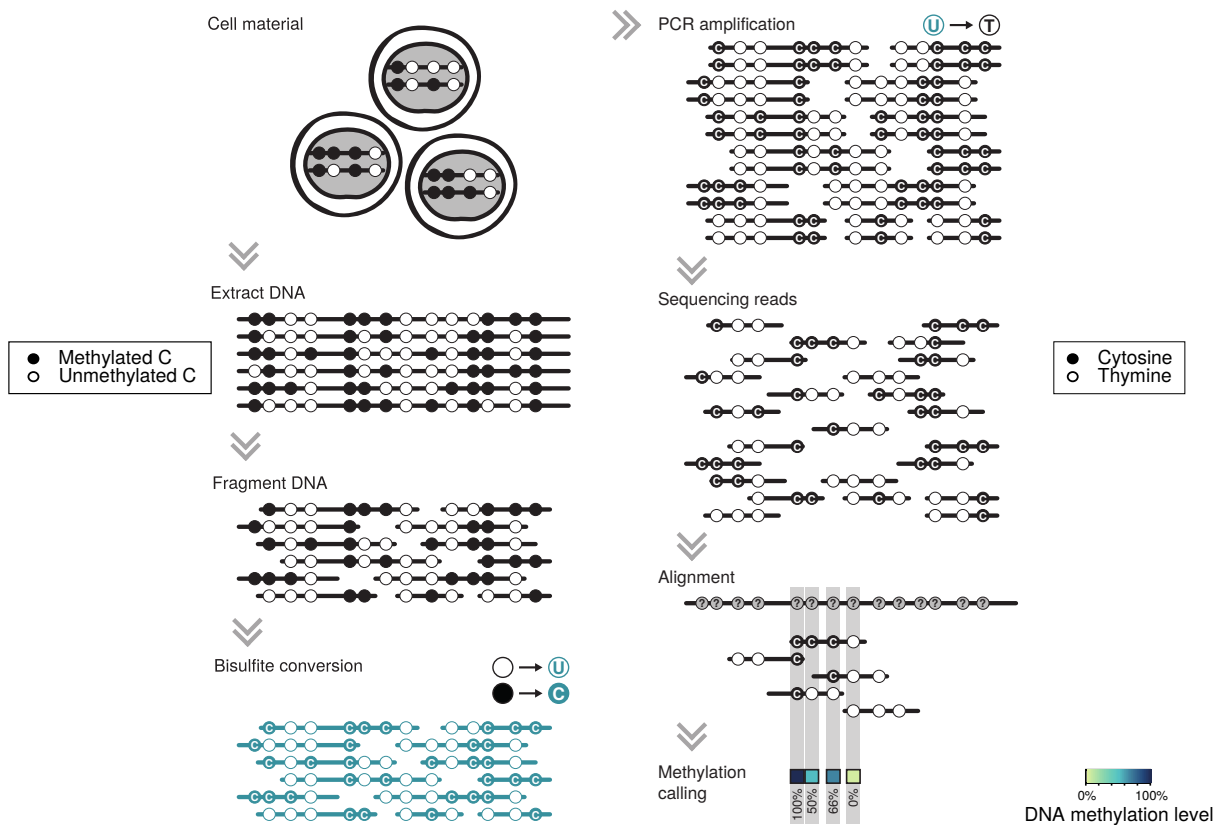


Figure 2.6: The workflow for genome-wide analysis of DNA methylation using WGBS comprises 7 steps: DNA extraction, DNA fragmentation, bisulfite conversion, PCR amplification, sequencing, alignment, and methylation calling. Modified from <https://doi.org/10.6084/m9.figshare.5285470.v1>, created by Fabian Müller.

of fragments that contain CpGs through enzymatic digestion with a methylation-insensitive restriction enzyme. In the original publication [104], the restriction enzyme *MspI* was used to cut the DNA at CCGG positions irrespective of the methylation state. Afterwards, a size selection is performed to enrich for fragments that are more likely to contain CpG dinucleotides. A library created using *MspI* assays mainly CpGs within CGIs. Since CGIs co-localize with gene promoters, the methylation states of CpGs assayed in a *MspI* library tend to be rather stable over different cellular states. Libraries created using an alternative restriction enzyme, *HaeIII* (restriction sequence GGCC), preferentially capture putative enhancer regions whose methylation states are more variable across different cellular states [105]. Multiple restriction enzymes can be combined in double-digestion libraries to combine the advantages of different enzymes [106]. Through enrichment of the fragments comprising CpG-dense regions, RRBS substantially reduces sequencing costs, since fewer sequencing reads have to be generated to cover 10–20% of the CpGs in the human genome. Furthermore, in contrast to the Illumina BeadArrays, RRBS can be applied to samples from any organism and is particularly useful for the analysis of murine samples, for which a microarray platform has only recently been proposed.

Enrichment-Based Approaches

In contrast to the size selection employed by RRBS, another approach for enriching sequencing fragments toward methylated fragments is methylated DNA immunoprecipitation sequencing (MeDIP-seq). This approach works analogously to ChIP-seq for genome-wide investigation of histone modifications. An antibody specific to 5mC is used to enrich for methylated fragments. The fragments can then either be hybridized to a microarray (MeDIP-chip [107]) or sequenced (MeDIP-seq [108]). Similarly, methyl-CpG-binding domain (MBD) proteins can be used to capture methylated sequences from the human genome, which can be subsequently sequenced. MBD-seq [109] and MeDIP-seq return relative enrichments of methylated fragments by comparing the sequencing reads generated from the selected fragments with the background, i.e., sequencing reads generated without prior immunoprecipitation or MBD-selection. Caveats of MBD-seq and MeDIP-seq include that co-methylation of adjacent CpGs is important for reliable precipitation and that fragments without CpGs may be precipitated by mistake. MBD-seq and MeDIP-seq return relative rather than absolute methylation levels, thus RRBS and WGBS are more frequently employed by the scientific community [110].

2.5 Basic Processing of DNA Methylation Data

Throughout this work, processed DNA methylation data will be utilized, i.e., data that can be represented in tabular form with CpGs as the rows and samples as the columns. However, data is not obtained in such a format from the sequencing machine or the microarray scanner. In this section, the steps necessary for obtaining processed DNA methylation data in tabular form based on the raw sequencing or microarray data will be presented. The focus is on data generated using the Illumina BeadArrays and on WGBS/RRBS data, and the section also presents crucial quality control and normalization steps.

2.5.1 BeadChip Arrays: From Intensity Data to a Data Matrix

DNA Methylation Microarrays

Intensity data is obtained by scanning the Illumina Methylation BeadArrays using an Illumina iScan or HiScan machine, which generate intensity data files (IDAT). These files are direct input to software packages such as *minfi* [111] or *methyllumi*⁴, which use IDAT files to obtain beta values from the red and green channels available as two separate files according to Equation 2.1. In addition to the raw intensity data, the Illumina manifest file maps internal identifiers of the microarray to genomic positions. Currently, this manifest file is only available for the human reference genome version ‘hg19’. *RnBeads*, which will be described in more details in Section 3.1, internally uses *methyllumi* for reading IDAT files. Data quality can be checked using the built-in control probes and further processed using CpG filtering and normalization methods.

Sample-Specific Quality Control In order to check for data quality of the Illumina BeadArrays, quality control probes for different processing steps have been established. These control

⁴<https://www.bioconductor.org/packages/release/bioc/html/methyllumi.html>

probes were constructed to exhibit high, medium, or low intensity values for successful experiments. Additionally, background control probes report if the overall signal intensity is substantially higher than expected. If several of the quality controls exhibit an unexpected distribution of intensity values for a particular sample, this sample should be excluded from downstream analysis. Similar to these control probes, the bead arrays comprise a few dozen highly variable SNP probes, which should show methylation values of 0, 0.5, and 1.0 for different genotypes. In a genetically matched setup, these probes can be used to check if the samples group together as expected. An unexpected clustering is an indication of a sample mix-up. The detection of sample mix-ups can be further supported by inferring additional sample properties such as donor sex or age from the methylation data (see Section 3.1).

CpG Filtering In addition to the removal of potentially unreliable samples, filtering of features (i.e., CpG sites) prior to the analysis of DNA methylation data is critical. Using the background control probes discussed earlier, a detection p-value can be computed for each CpG site individually. The detection p-value indicates if the detected signal is significantly different from the background signal [111], and sites with a high detection p-value (e.g., larger than 0.01) should be removed from the analysis. Such a removal step has been implemented in *RnBeads*' Greedycut algorithm, which iteratively removes CpGs from the analysis [112]. The Illumina BeadArrays comprise multiple CpG sites that are annotated to common SNPs with a minor allele frequency (MAF) in a general population higher than 1%. Using databases such as dbSNP [113], CpG sites located at these SNPs should be removed from downstream analysis since methylation differences detected at these positions can be due to genetic rather than epigenetic alterations. Similarly, CpGs located on the sex chromosomes are strongly different between the two sexes and are removed for most downstream analyses. Additionally, some CpG sites on the bead arrays have been shown to be cross-reactive, i.e., they are located in repetitive sequences or are highly homologous to other sequences. To avoid determining spurious associations, cross-reactive probes should be discarded [114, 91]. Most software packages for the analysis of Illumina BeadArrays contain preprocessing steps that comprise the steps mentioned above (see also Section 3.1).

Normalization Since the dynamic ranges of the two probe designs on the 450k and EPIC arrays are different, data normalization of the beta values is important. The respective normalization methods include quantile normalization, which match the distribution of type I probes to those of type II probes and *vice versa*. For instance, a beta-mixture quantile (BMIQ) normalization method has been introduced by Teschendorff et al. [115] for the 450k array and has also proven useful for the EPIC array. Additionally, the “dasen” normalization method from the *wateRmelon* R-package [116] accounts for known biases of EPIC array data. Since it has better runtime performance than BMIQ, it is the default normalization method used by *RnBeads*. A comprehensive, independent evaluation assessing both runtime performance and the effect of bias correction across different normalization methods is currently missing, and further normalization methods have been proposed, including SWAN [117], *functional normalization* [94], or *noob* [118].

Genotyping Microarrays

Calling Genotypes from Intensity Data To obtain genotype calls for the SNPs available on the Illumina genotyping microarrays (e.g., Infinium OmniExpress or Infinium OmniExpressExome BeadArrays), a genotyping algorithm is required. One such algorithm is *KRLMM*, which exhibits performance comparable with other genotyping algorithms [119]. In a first step, raw signal intensities obtained from the microarray are quantile normalized to adjust the distributions of the samples to one another. As the second normalization step, loess normalization is employed on the data. The genotyping process relies on k-means clustering of the normalized signal intensities, which assigns the samples for each of the SNPs into either one, two, or three categories (according to genotype AA, AB, and BB). To select the number of clusters for each SNP individually, the *KRLMM* algorithm uses a logistic regression classifier based on the residual sum of squares, the Mahalanobis distance, and the agreement with the Hardy-Weinberg equilibrium. The coefficients of the logistic regression classifier were trained on data from the HapMap project [120]. After selecting the number of clusters for each of the SNPs individually, k-means clustering is used for genotyping the samples.

Imputation Reference genotypes across many human populations have been created in projects such as the HapMap or the 1000 genomes project [121]. Thus, it suffices to assay a lower number of SNPs using microarrays in a study population and infer the genotypes of SNPs only present in the reference panel for the study at hand. This process, typically referred to as genotype imputation, first infers haplotypes from the data using, e.g., Hidden-Markov Models (HMMs). In the next step, missing SNP genotypes are inferred from the conditional probabilities of the model learned in the first step. Going into more details about the imputation procedure is beyond the scope of this thesis and further information can be found in the publications of the imputation methods *IMPUTE2* [122], *MaCH* [123], or the Michigan Imputation Server [124].

2.5.2 Sequencing-Based Approaches: Quality Control, Alignment, Quantification

In contrast to the Illumina BeadArrays, bisulfite sequencing data requires substantially more processing steps to produce reliable methylation calls of single CpGs across all the samples. Raw sequencing data is directly obtained from a sequencing machine. In a first step, raw binary base call (bcl) files are converted into read files (FASTQ format), which are human-readable files that comprise a single line of base calls for each sequencing read along with further quality information. Typically, multiple samples will be analyzed on the same lane of the sequencing flow cell, and FASTQ files need to be assigned to samples using an indexing strategy. This process, referred to as *demultiplexing*, generates a single FASTQ file for a single-read library and two FASTQ files (one for each read) for a paired-end library.

The steps necessary to generate single-CpG methylation calls from raw sequencing data are assembled into pipelines. International epigenomic consortia, such as the International Human Epigenome Consortium (IHEC [125]), BLUEPRINT [126], and the German Epigenome Program (DEEP), established such pipelines for routinely processing bisulfite sequencing data. Notably, the pipelines differ in the software tools they use and generate slightly different output formats. In the following, we present the steps and different software tools of such a pipeline, but do not present a concrete example of a pipeline. Such examples can be found, e.g., on the DEEP

GitHub page⁵.

Quality Control

Raw sequencing data is checked for data quality using metrics such as number of non-called bases (Ns), percentage of the different nucleotides, or number of PCR duplicates. An unusually high number of duplicates indicates an overamplification of the fragment during PCR. Unusual values of the quality statistics should be carefully investigated and the corresponding samples can potentially be excluded from downstream analysis. Tools such as *FastQC*⁶ give a comprehensive overview of various quality statistics and provide recommendations for assessing data quality. However, bisulfite treated reads require special investigation, since unmethylated cytosines are converted to thymines. Thus, only methylated cytosines will appear as cytosines in the sequencing reads, while methylated cytosines are heavily underrepresented throughout the genome. Consequently, the final sequencing reads show a low cytosine content in the quality control reports. Additionally, RRBS libraries show a substantially higher PCR duplication rate. In WGBS libraries, the start positions of reads are more or less randomly distributed across the genome and detecting the same start position of a read multiple times is unlikely. These sequences are marked as duplicates by *FastQC*, while RRBS reads are more likely to start at identical position due to the restriction enzyme treatment.

Trimming, Mapping, and Methylation Calling

After thoroughly checking for data quality, the sequencing data needs to be converted into single-CpG methylation levels. Sequencing adapters do not contain information about the sequence of interest and are removed from the sequencing files in the trimming step, using tools such as *TrimGalore!*⁷ or *cutadapt* [127]. Since FASTQ files contain raw sequence information and no information about the location of the signal, the sequence information of the fragments needs to be associated to a genomic region. This is achieved in the mapping step, where all sequencing reads are aligned to a reference genome at the position that best matches the read sequence. In comparison to aligning genomic reads, aligning bisulfite-converted sequences to a reference genome is substantially harder, since the DNA alphabet is virtually reduced to only three letters (A, G, and T, only very few Cs) with long T-stretches due to bisulfite conversion. While a thymine in the read can reflect both a thymine or an unmethylated cytosine in the original sequence, a thymine in the reference genome can never match a cytosine in the read. This problem is referred to as an asymmetric mapping problem. Current bisulfite read alignment tools such as *bsmap* [128], *bismark* [129], or *gemBS* [130] perform *in-silico* bisulfite conversion of the reference genome. The modified genome is then used as a reference for a genomic alignment tool such as *bowtie* [131]. *Bowtie* uses the Burrows-Wheeler transform to generate an index structure for matching the sequencing reads into the reference genome. The reads are aligned against the most recent version of the reference genome (currently 'GRCh38'/'hg38' for human). Further quality checks such as bisulfite conversion controls can be performed through spiked-in sequences with known DNA methylation states. After assigning the reads to a genomic position, the methylation level for each of the CpGs is computed using Equation 2.2 in

⁵<https://github.molgen.mpg.de/DEEP/comp-metadata>

⁶<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

⁷https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

the methylation calling step.

2.6 Methodological Background

I will now move from the biological motivation and background toward an introduction of basic mathematical definitions and computational methods that will be used throughout this work. Most of the definitions and concepts introduced below were obtained and modified from James, Witten, Hastie, and Tibshirani [132] and Hastie, Tibshirani, and Friedman [133]. The section requires a basic understanding of linear algebra.

2.6.1 Notations and Definitions

Throughout this work, data matrices representing methylation states of multiple samples in a population across multiple CpGs assayed will be used. We consider p as the number of rows in the matrix, which represent the features. Each row either comprises the methylation states of a single CpG dinucleotide or an aggregate value across a predefined genomic region, such as a promoter. n is the number of columns of the matrix and represents the observations/samples in the dataset. The data matrix D has dimension $p \times n$ and is the result of the DNA methylation mapping techniques and the basic processing steps mentioned above. Notably, each entry of D is in the interval $[0, 1]$.

To explore the data matrix and to generate hypotheses, various statistical learning techniques are applied to the data matrix. These techniques can be divided into *unsupervised* and *supervised* learning methods. While supervised learning aims to predict an outcome y given a set of observations (training data), unsupervised learning determines patterns and structures in the data without considering an output variable. If the output is a *categorical* variable, i.e., it has a finite number of possible values (mostly two), the supervised learning task is called *classification*, while *regression* predicts a *continuous* output variable, which can take on uncountably many values. If a model makes assumptions about the structure or distribution of the data, it is considered a *parametric* model and otherwise a *non-parametric* model.

In supervised learning, the objective is to minimize the *test error*, i.e., the error that the model makes when applied to a previously unseen dataset. The test error cannot easily be estimated during the training process, since an independent test dataset is often unavailable. In contrast, the *training error* is the difference between the model predictions and the true output in the training samples. It turns out that there are three components contributing to the test error according to the *bias-variance trade-off*. The *variance* describes the variability of the model using different training datasets. In contrast, the *bias* describes the systematic deviation of the model from the underlying relationship between input and output. The last component contributing to the test error is the *irreducible error*, i.e., the error or noise inherent to the data that cannot systematically be described.

2.6.2 Linear Regression

In the simplest regression model, known as linear regression, a linear model is constructed to describe the relationship between the input data matrix $D_{p \times n}$ and the output $y_{n \times 1}$. In a simple linear regression $p = 1$ and the goal is to minimize a *loss function*. This loss function describes the difference between the predictions \hat{y} and the true outputs y , where \hat{y} can be written as:

$$\hat{y} = \beta_0 + D^T \beta_1$$

where β_0 (intercept) and β_1 (slope) are coefficients to be estimated. However, p will typically be much larger than one and, within this thesis, even exceed the value of n .

If $p > 1$, the problem can be formulated as multiple linear regression. In the following, the scalar values β_0, \dots, β_p will be replaced by a vector β of dimension $(p + 1) \times 1$ and there will be an additional row in D for the intercept β_0 :

$$\hat{y} = D^T \beta$$

The goal of linear regression is to minimize the loss function. An example for a loss function, which is frequently used is the *Residual Sum of Squares* (RSS). The RSS describes the sum of squared differences between the predictions and the output vector for all samples:

$$\min_{\beta} \text{RSS}(\beta) = \min_{\beta} (y - D^T \beta)^T (y - D^T \beta)$$

It turns out that the solution to this optimization problem can be solved analytically and returns the *least squares* estimate. According to the *Gauss-Markov theorem*, the least squares estimate is, among all unbiased linear estimators, the one with the lowest variance, assuming that the true relationship between input and output is linear.

An important problem of least squares regression occurs when $p > n$, which is commonly the case in epigenomic research. In such a case, there is no unique solution of the minimization problem. To deal with this issue, the coefficient values in the optimization problem are regularized or a subset of the p features is selected. There is no unbiased linear estimator with a variance lower than the least squares estimate, but there might be *biased* estimators with substantially lower variance. The regularization term leads to a reduction of the variance of the estimated coefficients, i.e., the level of variability of the coefficient estimates obtained on different datasets is reduced. Two commonly used methods of regularized linear regression are *ridge regression* and the *Lasso* [134]. Both of the methods impose a penalty on the parameters to be estimated, which results in the modified optimization problems:

$$\text{Ridge regression: } \min_{\beta} (\text{RSS}(\beta) + \lambda \|\beta\|_2^2)$$

$$\text{Lasso: } \min_{\beta} (\text{RSS}(\beta) + \lambda \|\beta\|_1)$$

In contrast to ridge regression, the Lasso performs variable selection by setting some of the coefficients to exactly zero. The advantage of the regularization comes from the bias-variance trade-off. The regularization employed through ridge regression and the Lasso substantially decreases the variance, while introducing a bit of bias. In order to combine ridge regression and the Lasso, *elastic net regression* [135] was introduced, which combines the Lasso and ridge penalties using the hyperparameter α .

$$\min_{\beta} (\text{RSS}(\beta) + \lambda(\alpha \|\beta\|_2^2 + (1 - \alpha) \|\beta\|_1))$$

In contrast to linear least squares, this optimization problem does not have an analytical solution and is solved using coordinate descent methods [136]. To find optimal values for the hyperparameters α and λ , a *cross-validation* (CV) or *nested-cross-validation* scheme on a range of different user-defined values is recommended. Cross-validation is the process of (repeatedly) resampling training and disjoint test data from the dataset, training the model on the training dataset, and assessing performance using the test dataset.

In addition to the prediction task, linear models can be used to determine statistically significant associations between a pair of variables, such as gene expression values and DNA methylation. Due to their universal applicability, they will be used at various points throughout this thesis, especially in Chapter 3 and Chapter 5.

2.6.3 Logistic Regression

For predicting a categorical (here binary) rather than a quantitative output variable, classification methods can be used. *Logistic regression* is a classification method that uses the logistic function to map the output into the interval $[0, 1]$:

$$p = \frac{e^{d_i^T \beta}}{1 + e^{d_i^T \beta}} \quad (2.3)$$

where d_i is a column (observation) of D . Here, the output p can be interpreted as a probability for the observation d_i belonging to the first class, while the probability of belonging to the second class is $1 - p$. The goal of logistic regression is to obtain values for β such that the probabilities for the observations in the first class are maximal and minimal for the second class (i.e., $1 - p$ is maximal). In contrast to linear regression, logistic regression does not have an analytical solution, but can efficiently be solved using *maximum likelihood*. Briefly, maximum likelihood aims at finding the coefficients β that lead to a maximum value for equation Equation 2.3 for all data points belonging to the first class and maximizing $1 - p$ for all data points belonging to the second class. Using this notion, the likelihood is defined as the product of all probabilities (p values) for the observations in the first and $1 - p$ values for the observations in the second class. The maximum of this likelihood function (or equivalently the minimum of the negative, logarithmic likelihood) is obtained using the *Newton-Raphson* method.

Given the probability for each observation belonging to a class and the true class labels for each observation, a *confusion matrix* (Table 2.1) can be constructed:

Table 2.1: Possible outcomes of a binary classification task (confusion matrix).

		Predicted class	
		Positive	Negative
True class	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

From this matrix, *sensitivity* and *specificity* can be computed:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Both sensitivity and specificity depend on the threshold employed for differentiating between the two classes (e.g., the probability estimate returned by logistic regression). To visualize the values for sensitivity (y-axis) and 1-specificity (x-axis) for different thresholds, the *Receiver Operator Characteristic* (ROC) curve can be used. The *Area Under the Curve* (AUC) of the ROC curve is a measure for the quality of the binary classification task.

2.6.4 Matrix Decomposition

A typical data matrix D used within this thesis is high-dimensional (e.g., several hundred thousand rows and several hundred columns) and cannot be easily visualized. To overcome this problem, a special type of unsupervised learning tools called *dimension reduction* methods have been introduced. One such dimension reduction method, *principal component analysis* (PCA), operates on *eigenvalues* and a matrix of *eigenvectors*. To obtain eigenvalues and eigenvectors, *singular value decomposition* (SVD) of the centered data matrix D (row sums equal to zero) can be used:

$$D^T = USV^T$$

$$DD^T = (VS^T U^T)USV^T = V(S^T S)V^T$$

where U is an orthogonal matrix (assuming that D is real-valued), S comprises singular values in the diagonal (zeros elsewhere), and V is the matrix of eigenvectors of DD^T , which are also known as *principal component* directions of D^T . The principal components can be used for visualization of the data in low dimensional space. Principal component directions form an orthogonal basis by construction. Similarly, *Independent Component Analysis* (ICA) returns statistically independent components (to be discussed in Chapter 4). To allow for drawing the data in the paper or screen plane, the first two principal components are often used. Further, non-linear dimension reduction methods include t-distributed stochastic neighbor embedding (tSNE) [137] and Uniform Manifold Approximation and Projection (UMAP) [138], which are routinely used in single-cell data analysis.

We will use another matrix decomposition technique – non-negative matrix factorization (NMF) – for decomposing the data matrix into sources of variation, while imposing additional constraints specific to DNA methylation data. The process of decomposing the matrix into components of variation is called *deconvolution* and mimics restoring the original signal after it has been blurred by a filter in the *convolution* process. In the context of DNA methylation data, the convolution is imposed by a mixture of multiple cell types in a bulk tissue sample, as well as additional sources of variation (e.g., age, sex) that contribute to the measured DNA methylation signal in the data matrix. NMF will be introduced in detail in Chapter 4.

2.6.5 Clustering

Another instance of unsupervised statistical learning aims to group together samples that behave similarly with respect to the features. More specifically, the goal of *clustering* is to create groups (clusters) of observations that are more similar to the observations inside the group than to those outside of the group. Two frequently used clustering methods are *hierarchical clustering* and *k-means clustering*. Both clustering approaches employ a *distance metric* to define a dissimilarity between two data points, and the Euclidean distance is commonly used:

$$\|x - y\|_2$$

where x, y are two observations (p -dimensional vectors). Additional distance metrics include the correlation-based distance (1 - (Pearson) correlation of the vectors), Manhattan distance, and Mahalanobis distance.

Hierarchical Clustering Hierarchical clustering of the samples can be performed by a simple algorithm. The algorithm initializes each sample as a cluster on its own (a singleton) and iteratively merges the two most similar (least dissimilar) clusters, until only a single cluster is left (bottom-up clustering). To define the dissimilarity between two clusters, different *linkage* criteria have been introduced: average, complete, and single linkage. Complete and single linkage use the maximum and minimum distance, respectively, between any two data points in the two clusters to be compared, while average linkage uses the average dissimilarity between all the pairs of data points in the two clusters:

$$t(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} t_{ij}$$

where G, H are the clusters, N_G, N_H the number of observations per cluster, and t_{ij} the dissimilarity between data points i and j . Using this approach, a *dendrogram* is generated as a tree structure that represents the merging process and a clustering is obtained by a horizontal cut of the dendrogram. Dendrograms will be used along with *heatmaps* throughout this thesis, where both rows and columns of a matrix are hierarchically clustered. In contrast to hierarchical clustering, *k-means* clustering requires the number of clusters to be defined *a priori*. Details about the *k-means* clustering algorithm can be found in the literature [132].

Louvain Clustering Louvain clustering is a computational method that uses graph concepts to define clusters [139]. From the set of observations and an associated distance metric (e.g., Euclidean distance, correlation-based distance), a weighted graph can be constructed. In this graph, each node is an observation and the weight of an edge (w) corresponds to the similarity (1-distance) between the two data points. Given this graph and a partition into clusters, the *modularity* Q can be defined as:

$$Q = \frac{1}{2m} \sum_{i,j} (w_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

where i, j are data points (nodes), w_{ij} is the weight of the edge between i and j , $k_i = \sum_j w_{ij}$ is the sum of edge weights to all other nodes from node i , c_i is the cluster to which node i belongs to, $\delta(c_i, c_j)$ is one if i and j belong to the same cluster and zero otherwise, and $m = \frac{1}{2} \sum_{i,j} w_{ij}$ [139]. Intuitively, Q will be large if for all clusters the sum of edge weights between nodes within the same cluster (w_{ij}) is large in comparison to the product of all edge weights (k_i, k_j). Thus, partitions of the graphs with high values for Q have many connections within the nodes of a cluster, but only few connections to the nodes outside of the cluster.

The Louvain clustering algorithm obtains a clustering of the network that maximizes the modularity Q using an iterative, two-stage procedure. First, each cluster comprises a single observation similar to hierarchical clustering. The two stages of the algorithm are as follows:

1. For each node i in the graph, investigate whether an increase of modularity can be achieved by placing i into the cluster of any of its neighbors j . The node is placed in the cluster resulting in the largest increase in modularity. The first stage is executed sequentially and repeatedly until no further increase of modularity can be achieved.
2. Construct a new network, where each node is a cluster from stage 1, and the edges between two clusters are the sum of weights of the edges between nodes in the two clusters. Self-loops represent the edges between nodes in the same cluster. Apply stage 1 to this new graph.

This whole procedure is iterated, until no further improvement of the modularity can be achieved [139]. The final stage of the algorithm returns clusters of data points, and Louvain clustering does not require the number of clusters to be defined *a priori*.

DNA Methylation Heterogeneity Between Phenotypes

In this chapter, I will present two projects: First, I will show how we can use epigenomic differences between groups defined by a phenotype to gain insight into epigenetic regulation and its association with diseases. To facilitate the analysis of between-group heterogeneity, DNA methylation data can be comprehensively analyzed with the software suite RnBeads. Together with the original developers of RnBeads (Yassen Assenov, Fabian Müller, and Paolo Lutsik), and with Christoph Bock as the main supervisor, I developed an updated version of RnBeads. I contributed new state-of-the-art modules (age and sex prediction, missing value imputation, differential variability, segmentation), applied the updated package to cancer samples, and performed a benchmark of RnBeads in comparison to other published tools. The first part of the chapter is a modified version of Müller et al. [140] published in Genome Biology (2019), in which we propose RnBeads 2.0. Additionally, I discuss the relationship between DNA methylation and the human aging process, which is follow-up work based on my Master's thesis (Scherer [141], 2016). In a joint project with Lisa Eisenberg (née Handl) and Nico Pfeifer, we investigated the applicability of unsupervised domain adaptation for epigenetic age prediction. I contributed processed DNA methylation data as input to the model and provided the baseline comparison model.

In the second part of this chapter, I investigate the relationship between donor genotype and DNA methylation patterns. Methylation quantitative trait loci (methQTLs) are genetic alterations (SNPs) that correlate with the DNA methylation state of individual CpGs. An important biological question is whether those interactions are shared across different cell types or whether they are cell-type specific. To answer this question, I developed a novel software tool (MAGAR) and used established computational frameworks to jointly analyze DNA methylation and genotyping data from four tissues assayed in the context of the SYSCID¹ project. The research was performed in close collaboration with Gilles Gasparoni (Genetics Department), Souad Rahmouni and Michel Georges from the university of Liège, Paul Lyons from Cambridge university (UK), Yurii Aulchenko and Tatiana Shashkova, and Jörn Walter as the main supervisor. A manuscript describing the software tool and the analysis of methQTLs is currently in preparation.

¹<http://syscid.eu/>

3.1 *RnBeads* 2.0: Comprehensive Analysis of DNA Methylation Data

3.1.1 Overview of DNA Methylation Analysis Tools

In order to guide diagnosis and therapy selection in a clinical setting, biomarkers for a disease are required. Such a biomarker can provide information about the presence of the disease, the subtype of the disease, or is informative about patient prognosis. DNA methylation has recently emerged as a premier candidate for the discovery of epigenetic biomarkers [142, 70] due to the increasing availability of DNA methylation data. To detect novel DNA methylation-based biomarkers, a group of samples obtained from affected individuals is compared to a control group comprising healthy individuals. Additionally, different subgroups within the group of affected individuals can be defined and used for discovering biomarkers of disease subtypes or patient prognosis. No control group is required for such an analysis (see Section 3.1.3). In brief, the DNA methylation patterns of the groups are compared, and CpGs or genomic regions identified that are significantly different between the groups. This type of analysis, referred to as an EWAS, requires extensive data processing to focus on a set of reliably detected CpGs and necessitates additional exploratory analysis to generate and confirm hypotheses.

Using the EWAS methodology, DNA methylation aberrations have been associated with various diseases (cf. Section 2.3 [75, 143]). Notably, the functional implications of DNA methylation aberrations remain largely unknown and need to be addressed, e.g., through functional assays such as CRISPR/Cas9. The typical input to EWAS is DNA methylation data generated through bisulfite sequencing or microarray technology. Further information about the samples such as tissue or cell type, phenotypic data (donor age, sex, etc.), and sample grouping (e.g., disease versus control) is required. A bioinformatic workflow for EWAS comprises the following steps: (i) data import, (ii) quality control, (iii) identification of a set of reliable CpGs across the samples and removal of technical biases (iv) exploratory data analysis, and (v) association of DNA methylation heterogeneity with sample annotations (differential analysis). Most bioinformatic tools support individual steps of this workflow (reviewed in [144, 145, 146, 147] and summarized as a feature table in Supplementary Table A.1), while integrative tools are still scarce.

The *RnBeads* software package [112] is an R/Bioconductor package providing a pipeline for start-to-finish analysis of DNA methylation data. Notably, *RnBeads* supports both bisulfite sequencing and microarray data and allows for data integration across different technologies. It follows the standards and practices established by epigenomic consortia, such as the International Human Epigenome Consortium (IHEC). Since its original release in 2012 and the initial publication in 2014, *RnBeads* has become a well-established software tool. However, it is necessary to routinely update software packages by implementing novel analysis strategies and to continually improve the software. During this work, we extended the original software package with state-of-the-art analysis methods motivated by user feedback and feature requests and improved computational efficiency. Novel features include epigenetic age prediction, improved support of missing values, and analysis of differential variability. The updates have been collected in a new release of the software package (*RnBeads* 2.0) and position *RnBeads* as an integral tool in many standardized DNA methylation analysis workflows. *RnBeads* is actively used in the context of the German Network for Bioinformatics Infrastructure (de.NBI²)

²<https://www.denbi.de>

and the EU-funded SYSCID project³. We present an use case of the updated software package by analyzing a childhood bone cancer cohort (Ewing sarcoma) and find strong indications of tumor and DNA methylation heterogeneity in stem cells. Additionally, we present a benchmark of *RnBeads*2.0 in comparison to other software tools for the analysis of DNA methylation data.

3.1.2 Analyzing DNA Methylation Data with *RnBeads*

RnBeads Overview

RnBeads is a modular pipeline for the analysis of DNA methylation data that comprises seven core modules: data import, quality control, preprocessing (i.e., filtering and normalization), tracks and tables (i.e., export of processed data for visualization in a genome browser), covariate inference (e.g., predicting epigenetic age and cell-type composition), exploratory analysis (e.g., dimension reduction, global distribution of DNA methylation levels, hierarchical clustering), and differential analysis between two user-defined sample groups (Figure 3.1). For each of the analysis modules, *RnBeads* creates an interactive HTML report describing the analysis that has been conducted and provides associated plots. These reports can be used to share an analysis within the scientific community and facilitate the reproducibility of the results.

The core structure has already been developed for the original publication in 2014, but the modules have been extensively revised and extended in the updated version. Specifically, *RnBeads* 2.0 provides new/extended functionality for:

1. **Support for additional data types:** *RnBeads* now also supports data generated by the newest version of the Illumina BeadArray series, the EPIC array, and allows for the integration of different data types (Illumina BeadArrays and RRBS/WGBS) into a combined analysis. This extended functionality facilitates integrative analysis of datasets generated by different epigenomic consortia or by different technological platforms. CpG-wise methylation calls obtained from different technologies are mapped to overlapping genomic locations and aggregated into a combined data matrix.
2. **Additional analysis and inference algorithms:** *RnBeads* now handles missing values in the DNA methylation data matrix (missing value imputation). To obtain estimates of a sample's epigenetic age, we incorporated DNA methylation-based age prediction in a platform-aware manner for both array-based and sequencing-based datasets (see Section 3.2). These estimates can be useful for correlating accelerated epigenetic aging to physiological states [148] or to detect potential sample mix-ups. We also incorporated estimation of the overall immune cell content of a sample using the LUMP algorithm [149], which is motivated by immune infiltration into tumors and provides a generic estimate of tumor purity. New methods for quantifying DNA methylation variability [150, 151], i.e., detecting differentially variable CpGs (DVCs), contribute to the detection of epigenetic risk loci for cancer predisposition, and region set enrichment analysis using the LOLA tool [152] facilitates the interpretation of DMCs/DVCs. Additionally, *RnBeads* 2.0 supports DNA methylation-based segmentation through the *MethylSeekR* approach [153].

³<https://syscid.eu>

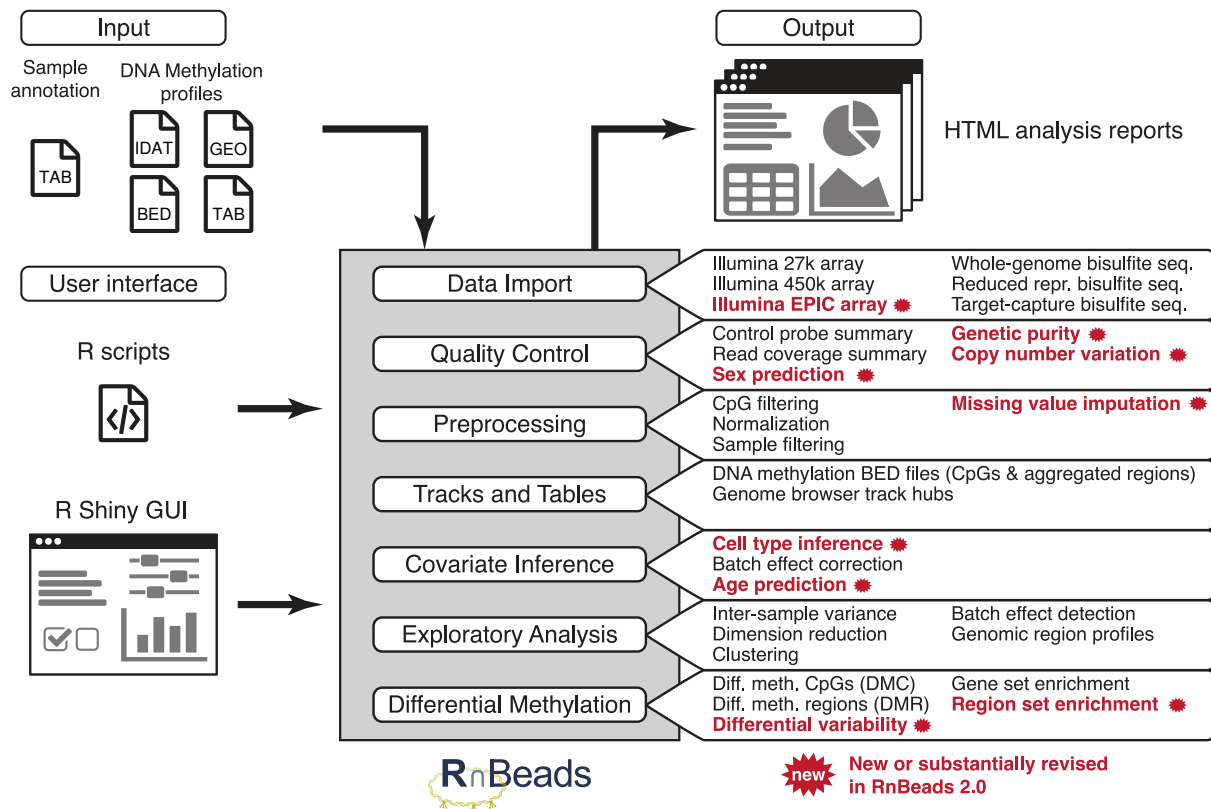


Figure 3.1: The *RnBeads* analysis workflow comprises seven modules for DNA methylation data analysis. *RnBeads* supports various kinds of DNA methylation data as input and requires phenotypic information as a sample annotation sheet. An *RnBeads* analysis can be configured using R scripting or the R/Shiny-based GUI *RnBeadsDJ*. Each of the analysis modules creates an HTML report describing the analysis steps performed (described on the right side) and associated results. TAB, tabular file; BED, Browser extensible data (RRBS/WGBS data); GEO, download from the Gene Expression Omnibus (GEO) repository

- Graphical user interface (GUI):** In order to make *RnBeads* more readily accessible also for non-bioinformaticians and for users with limited R/Bioconductor knowledge, we created a new R/Shiny-based GUI to configure and execute an *RnBeads* analysis. Along with the HTML reports facilitating the exploration and distribution of the results, this new interface improves usability.
- Computational efficiency:** *RnBeads* supports parallelization and allows for automatic distribution of jobs across the nodes of a high performance computing (HPC) cluster, now supporting two of the most widely-used job scheduling systems, the Sun Grid Engine (SGE⁴) and the Simple Linux Utility for Resource Management (SLURM⁵).

Using *RnBeads* 2.0, we analyzed hundreds of RRBS samples obtained from Ewing sarcoma patients [154] on an HPC cluster. We discuss some of the feature updates in more detail below and focus on the Ewing sarcoma dataset as a use case dissecting tumor heterogeneity. Lastly, we present a benchmark of *RnBeads* in comparison to other software tools.

⁴<https://docs.oracle.com/cd/E19279-01/820-3257-12/n1ge.html>

⁵<https://slurm.schedmd.com/documentation.html>

Dataset Description

Tumor heterogeneity is a key property of many cancers that is present at multiple molecular layers including the epigenetic layer [155]. Comprehensively assessing epigenetic heterogeneity is especially critical for bulk tumor samples, since different cell types (e.g., cancer cells, immune cells) in the sample contribute to the overall signal. We analyzed a recently published cohort of the childhood cancer Ewing sarcoma, with 188 samples comprising Ewing tissue samples, Ewing cell lines, and mesenchymal stem cells (MSC) obtained from healthy donors and from Ewing sarcoma patients (eMSCs, GEO accession: GSE88826 [154]). The pediatric bone cancer Ewing sarcoma exhibits low genetic, but substantially elevated epigenetic heterogeneity [154, 156]. We used *RnBeads*' novel differential variability module (details below) to investigate tumor and stem cell heterogeneity.

Notably, there are multiple methods for mapping DNA methylation data genome-wide (Chapter 2), and each of the methods delivers information on DNA methylation for a set of partially overlapping CpGs. Since different epigenomic consortia, such as DEEP or IHEC, generated epigenomic data using different technologies, there is a need for a software tool that integrates DNA methylation data across different technologies. Most of the available software tools focus either on microarray or bisulfite sequencing data (Supplementary Table A.1), and a computational suite for joint analysis is required. *RnBeads* handles any dataset providing single-CpG resolution and is capable of integrating different genome-wide assays, including WGBS, RRBS, and DNA methylation microarrays (27k, 450k, EPIC). Additionally, *RnBeads* supports enrichment-based assays (e.g., MeDIP-seq, MBD-seq) given that their relative output (read enrichment) has been converted into absolute, single-CpG methylation calls using available bioinformatic tools [157, 108]. *RnBeads* relies on pre-compiled annotation packages, which have been generated for different versions of the human ('hg19', 'hg38') and the mouse genome ('mm9', 'mm10'), as well as the rat reference genome ('rn5'). Custom annotations for other species can be generated using the *RnBeadsAnnotationCreator* package⁶.

Computational Scalability

Since DNA methylation data matrices are usually high-dimensional, they potentially do not fit into the random access memory (RAM) of a standard working station. Thus, *RnBeads* stores large matrices on disk rather than in main memory using the *ff* R-package⁷. The *foreach*⁸ and *doParallel* packages⁹ allow for different jobs to be automatically distributed across different cores of a machine. We have also implemented an interface that automatically distributes parts of the *RnBeads* analysis across the nodes of an HPC cluster. In addition to SGE, *RnBeads* supports the SLURM job scheduling system. This allows for *RnBeads* to be executed in the de.NBI cloud¹⁰. With the flexible option settings available in *RnBeads*, time-consuming or memory-consuming steps of the analysis pipeline can be deactivated and we provide pre-defined option settings for different computational configurations. All these features enable analysis of hundreds of bisulfite sequencing samples or thousands of Infinium microarray samples in a single execution of the pipeline.

⁶<https://rnbeads.org/tutorial.html>

⁷<https://CRAN.R-project.org/package=ff>

⁸<https://CRAN.R-project.org/package=foreach>

⁹<https://CRAN.R-project.org/package=doParallel>

¹⁰<https://cloud.denbi.de/>

Tool Comparison

In a review paper [158], the following tools for the analysis of DNA methylation data generated with microarrays have been evaluated: *minfi* [111], *methylumi* [159], *wateRmelon* [116], and *ChAMP* [160]. We compared runtime performance and peak memory consumption of *RnBeads* to these software packages. With *methyKit* [161], we also included a package supporting the analysis of bisulfite sequencing datasets into the benchmark. For evaluation of the tools focusing on microarray data, we used a dataset comprising peripheral blood samples from 732 healthy individuals [162] and benchmarked the performance on bisulfite sequencing data using a mouse RRBS dataset (GEO accession number GSE45361, 6 adrenal gland and 11 liver samples [163]) and a human WGBS dataset (12 hepatocyte samples) from the DEEP project. Thus, we covered the typical use cases of DNA methylation data analysis. All benchmarking runs were executed on a Debian Wheezy machine (32 cores@1.2 GHz, 126 GB RAM, R-version 3.5.0).

Since most of the tools provide different parameter settings for conducting different depths of analysis, we benchmarked three parameter settings separately (Table 3.1): (i) data import, (ii) core modules, and (iii) most features enabled (comprehensive analysis). Each of the settings was tested in three independent executions. Furthermore, we comprehensively evaluated and listed features available in different tools for the analysis of DNA methylation data in a table (Supplementary Table A.1). We included those Bioconductor packages for DNA methylation analysis that are widely used in the scientific community according to the Bioconductor download statistics. Tools were selected that provide more than an individual task of the data analysis such as data handling or normalization. Additional tools outside of the Bioconductor ecosystem were selected based on literature review.

Details on *RnBeads* Extensions

Missing Value Imputation Missing values in DNA methylation datasets are a recurring issue and constitute an important analytical challenge. For microarray data, missing values can arise from masking values with high detection p-values due to insufficient signal intensities. They arise from insufficient read coverage in WGBS data. Especially for RRBS, where the sites selected for sequencing are influenced by the restriction enzyme digestion, the number of missing values when combining many samples into a single data matrix can be substantial. Thus, *RnBeads* provides different solutions including means and medians across samples or CpGs, random sampling from other samples in the dataset, and k-nearest neighbors (KNN) imputation. Although KNN imputation has originally been developed for gene expression microarrays [164], it also has successfully been applied to DNA methylation data [44, 83]. If a sufficient number of nearby points are available, KNN estimations are well suited for replacing missing values and should be favored for microarray-based datasets. The mean and median imputation approaches have been implemented especially for the analysis of datasets comprising high numbers of missing values such as some low-coverage bisulfite sequencing datasets. *RnBeads* leaves it to the user which, if any, of the missing value imputation methods is employed and also supports removal of all sites that contain any missing value across the samples. The choice of imputation method can be critical and should be carefully considered in the analysis setup.

Table 3.1: Parameter settings used for benchmarking tools for DNA methylation analysis. n.a.= not available

Software version	Data import only			Core modules			Comprehensive setting		
	450k	RRBS	WGBS	450k	RRBS	WGBS	450k	RRBS	WGBS
<i>RnBeads</i> version 2.0	Import	Import	Import	Import, Normalization, Clustering, DMC+DMR calling	Import, Normalization, Clustering, DMC+DMR calling	Import, Normalization, Clustering, DMC+DMR calling	rnb.run.analysis	rnb.run.analysis	rnb.run.analysis
<i>minfi</i> version 1.12.2	Import	<n.a.>	<n.a.>	Import, QC, Preprocessing	<n.a.>	<n.a.>	Import, QC, Preprocessing, DMC calling	<n.a.>	<n.a.>
<i>methylumi</i> version 2.26.0	Import	<n.a.>	<n.a.>	Import, Normalization, SVD, DMC calling	<n.a.>	<n.a.>	<no additional analyses modules provided>	<n.a.>	<n.a.>
<i>ChAMP</i> version 2.10.1	Import	<n.a.>	<n.a.>	Import, Normalization, SVD, DMC+DMR calling	<n.a.>	<n.a.>	champ.process (ComBat disabled; QCplots reduced to 'mdsPlot' and 'densityPlot')	<n.a.>	<n.a.>
<i>wateRmelon</i> version 1.24.0	Import	<n.a.>	<n.a.>	Import, QC, Preprocessing, Testing for sex specific sites	<n.a.>	<n.a.>	Import, QC, Preprocessing, Testing for sex specific CpGs, Estimation of cell-type composition	<n.a.>	<n.a.>
<i>methyKit</i> version 1.6.1	<n.a.>	Import	Import	<n.a.>	Import, Filtering, Clustering, Batch correction, DMC calling	Import, Filtering, Clustering, Batch correction, DMC calling	<n.a.>	<no additional analyses modules provided>	<no additional analyses modules provided>

Table 3.2: Bisulfite sequencing datasets used for the development of a sex classifier.

Dataset	Source	Organism	Assay	#Samples	♀	♂
BLUEPRINT	August 2016 release	human	WGBS	188	99	89
DEEP data	version 11/2016	human	WGBS	31	14	17
Kiel Cohort	Szymczak et al. [165]	human	RRBS	239	106	133
Ewing tissue samples	Sheffield et al. [154]	human	RRBS	96	53	42
Reizel GEO	Reizel et al. [166]	mouse	RRBS	152	94	58
				706	366	339

Sex Prediction Sample sex is valuable information that should be included as a confounding factor in most epigenomic studies including differential analysis of DNA methylation. Patient sex can be reliably predicted using the relative signal intensities obtained for the sex chromosomes compared to the signals of the autosomes. *RnBeads* uses the average signal intensities of the microarray probes on the X- and Y-chromosome in comparison to the autosomes as input to a logistic regression model. The output of the model is a probability for the biological sex of the sample. Similarly, for bisulfite sequencing data, *RnBeads* quantifies the sequencing coverage for the sex chromosomes in comparison to the coverage for autosomes and provides a pre-trained logistic regression model. The newly developed bisulfite sequencing classifier has been trained and validated on a large dataset, comprising both human and mouse samples (Table 3.2, cross-validation accuracy: 94.3%). Since data obtained on rat samples is scarce, no robust classifier could be trained for rat. Predicted sex can be used to fill in missing annotations in the sample sheet or be used as a quality control tool to reveal potential sample mix-ups (cf. Section 4.1.3). We would like to point out that a deviation of the predicted sex from the annotated sex is merely an indication for a potential problem with the sample that should be further investigated.

Differential Variability Analysis When comparing different groups of samples, e.g., control samples versus a group of cases, the groups cannot only differ in terms of their average DNA methylation level, but also in the DNA methylation variability within the groups. Most prominently, DNA methylation profiles (i.e., different samples) obtained from different cancer patients can be substantially more variable than the samples of a control cohort. Differential variability methods have been introduced for determining which of the CpGs are affected by differences in DNA methylation variance. *RnBeads* supports two algorithms for quantifying differential variability between two groups of samples: *diffVar* [150] and *iEVORA* [151]. To test for differences of the variances within the two groups, *diffVar* employs an empirical Bayes framework and *iEVORA* uses the Bartlett test. The Bartlett test is a statistical test that assesses whether the variances of two samples are significantly different. Notably, *diffVar* allows for additional covariates to be considered and thus enables accounting for confounding factors such as age and sex, similar to *RnBeads*' differential methylation analysis. *RnBeads*' differential variability module follows closely the structure implemented in its differential methylation module for identification of DMCs and DMRs (see [112] for details). To that end, a ranking scheme is employed based on three statistics:

1. the (false discovery rate (FDR)-)adjusted p-value of either *diffVar* or *iEVORA*

2. the difference of the group-wise variances
3. the log-ratio of variances in the two groups

Each of these scores is ranked individually, and the final rank of a CpG site is the worst (i.e., highest) rank of the three scores. Thus, *RnBeads* uses the p-value from a statistical test and estimates of the effect size (variance difference and log-ratio of variances). Similar to the differential methylation module, *RnBeads* provides an automatically-generated rank cutoff to select those sites with the strongest effect (see documentation of function `auto.select.rank.cut` in the *RnBeads* manual for details¹¹). Alternatively, the user can select a defined number of best-ranking sites. The user can also follow a more classical p-value cutoff scheme. The computed statistics are aggregated over pre-defined genomic regions, including gene bodies, promoters, CpG islands, or custom region annotations. Notably, *RnBeads* reports all CpGs/regions as DVCs/DVRs and leaves it to the user to either select a rank cutoff or to define a p-value cutoff.

DNA Methylation-Based Segmentation Genome-wide patterns of DNA methylation are organized into broader domains in accordance with the organization of chromatin into euchromatin and heterochromatin. Partially methylated domains (PMDs) are regions in the human genome that exhibit variable DNA methylation patterns with an overall lower methylation level than the genomic average. They comprise up to 75% of the genome and are largely cell-type specific [18]. Genomic regions that are not classified as PMDs can be further subdivided into highly methylated domains (HMDs), lowly methylated regions (LMRs), and unmethylated regions (UMRs) according to their average methylation level and CpG density. *MethylSeekR* is a two-stage software tool for genome-wide segmentation of WGBS data into PMDs, HMDs, LMRs, and UMRs [153]. In the first step, a hidden-markov model (HMM) is employed for segmenting the genome into PMDs and non-PMDs using parameters of the DNA methylation distribution estimated from the observed read counts in fixed-sized windows of 101 CpGs. This number of CpGs has been selected as a reasonable default value in the original publication [153], but is available as a tool parameter. Non-PMDs are further subdivided into HMDs, LMRs, and UMRs using a rule-based workflow. We integrated genome-wide segmentation using *MethylSeekR* into *RnBeads* and support all sequencing-based assays, including WGBS and RRBS. The integration is based on a script that has been kindly provided by Abdulrahman Salhab from the Genetics/Epigenetics department.

Furthermore, we developed an extension of the *MethylSeekR* approach that supports datasets produced using the EPIC array. Instead of using the fixed-sized window as in the original approach, we used a KNN technique according to the genomic distance between CpGs to estimate the parameters as input to the HMM. In contrast to the PCA-based segmentation approach implemented in the *minfi* [111] R-package, which also allows for classification into PMDs/non-PMDs, we support segmentation using individual samples rather than biological or technical replicates. Using the new approach, we found reasonable concordance of segmentations based on EPIC data in comparison to WGBS data from matched samples [167] (unpublished work together with Malte Groß).

¹¹<http://bioconductor.org/packages/release/bioc/manuals/RnBeads/man/RnBeads.pdf>

3.1.3 Application of *RnBeads* in Cancer and Comparison to Additional Software Packages

We validated and showcased the new features in four datasets available on the *RnBeads* website¹² and will specifically highlight the new differential variability module by investigating DNA methylation dynamics in a childhood bone cancer cohort in this work.

Quantifying DNA Methylation Heterogeneity in Ewing Sarcoma

We used *RnBeads* to process and analyze 188 samples obtained from Ewing sarcoma patients that have been assayed using RRBS. To focus on a set of highly reliable CpG sites, we retained those sites with sequencing read coverage of five or more in at least 50% of the samples. This resulted in a final set of 2,217,786 CpGs, which were further aggregated across putative regulatory elements defined by the Ensembl Regulatory Build [8]. Using PCA, we found the expected separation of samples into Ewing tissue samples, Ewing cell lines, and MSCs, with substantially higher between-sample heterogeneity in the Ewing tissue and Ewing cell line groups (Figure 3.2A). In a differential analysis, we used *RnBeads*' differential methylation and differential variability modules to compare primary tumors with the cell lines. Strikingly, most of the DMRs had higher average DNA methylation levels in the cell lines (Figure 3.2B) and were hypomethylated in the tissue samples. Furthermore, elevated variance was observed in the cell lines (Figure 3.2C).

To biologically interpret the detected differences between the primary tumors and the cell lines, we conducted LOLA enrichment analysis [152] on the DMRs and DVRs. We found different enrichments for DMRs/DVRs indicating that differential methylation and differential variability analysis provide complementary information on the DNA methylation landscape (Figure 3.2D-F). Hypermethylated DMRs in Ewing sarcoma cell lines were preferentially located in DNaseI-hypersensitive sites identified in various tissue samples obtained from healthy individuals (Figure 3.2D). This observation is an indication of widespread silencing of non-essential regulatory regions in cell lines. On the other hand, hypervariable regions were enriched for TFBS and histone modifications specific to cancer cell lines and ESCs (Figure 3.2F). This indicated that the Ewing sarcoma cell lines showed elevated regulatory plasticity compared to the primary tumors.

Differential methylation and differential variability analysis can be used to characterize the DNA methylation landscape in association to a disease and thus to quantify DNA methylation heterogeneity. *RnBeads* provides a comprehensive list of functions to analyze DNA methylation data and can be used to analyze RRBS data after region-based aggregation of single CpG values across pre-defined regulatory regions.

Comparison to other Software Tools for DNA Methylation Analysis

Computational runtime and requirements for computational infrastructure can be a main bottleneck for epigenomic data analysis. Thus, we compared the computational efficiency of *RnBeads* with other software tools for the analysis of DNA methylation data including *minfi*, *wateRmelon*, *methylumi*, and *ChAMP* [111, 116, 160] for microarray data. Similarly, we compared *RnBeads* to *methyKit* [161] both on RRBS and WGBS data (see Table 3.1). The three scenarios we

¹²<https://rnbeads.org/methylomes.html>

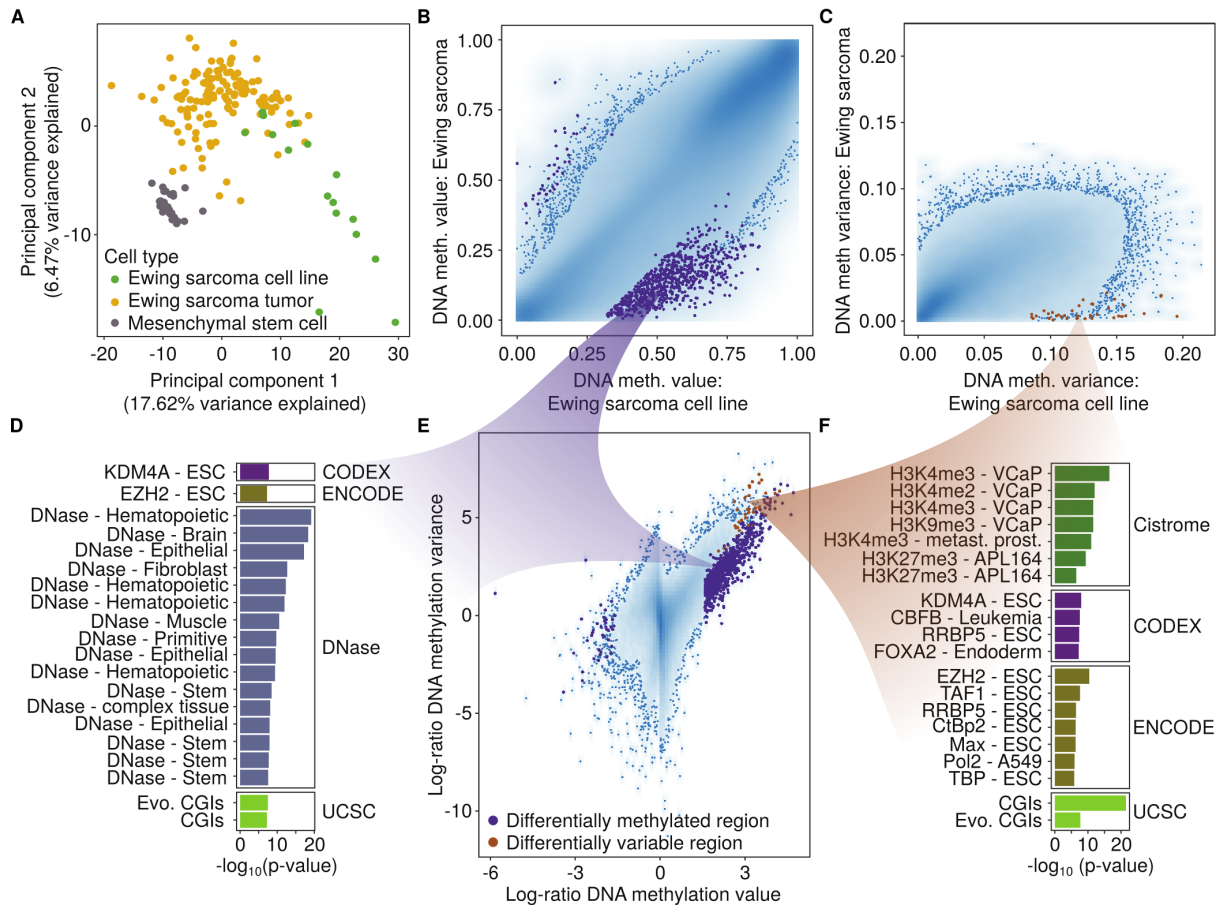


Figure 3.2: DNA methylation heterogeneity in the childhood cancer Ewing sarcoma. **A:** PCA of the RRBS samples comprising Ewing sarcoma tumors, cell lines, and MSCs. DNA methylation values were aggregated across putative regulatory regions. **B:** Scatterplot comparing the average DNA methylation levels per putative regulatory region between Ewing sarcoma tumors (N=140) and Ewing sarcoma cell lines (N=16). Marked in purple are those regions that had a differential DNA methylation rank lower than the automatically selected rank cutoff. **C:** Scatterplot comparing DNA methylation variability between Ewing sarcoma tumors and cell lines. Significant DVRs are marked in brown. **D:** LOLA enrichment analysis for the DMRs in panel **D** and in panel **E**. Visualized are the negative common logarithms of the enrichment p-values for different region databases. **E:** Scatterplot comparing the log-ratios between mean DNA methylation level and variance in Ewing sarcoma tumors and cell lines. Points are colored according to the definitions in **B** and **C**. **F:** LOLA enrichment analysis for DMRs shown in panel **C** and in panel **E**. ENCODE, transcription factor binding sites ChIP-seq profiles from the Encode [168] project; CODEX, ChIP-seq profiles from the Codex database [169]; Cistrome, ChIP-seq profiles from the Cistrome project [170]; DNase, DNaseI-hypersensitive sites; UCSC, annotations obtained from the UCSC genome browser [171].

investigated were: (i) data import, (ii) core modules, (iii) comprehensive analysis (Figure 3.3), to be able to compare different depths of analysis. Notably, *RnBeads* is the only tool that supports both microarray-based and bisulfite sequencing-based analyses.

The tools *minfi*, *methylumi*, and *wateRmelon* only provide a basic set of analysis options, but were faster than *ChAMP* and *RnBeads*, since the latter two need to prepare a more complex data

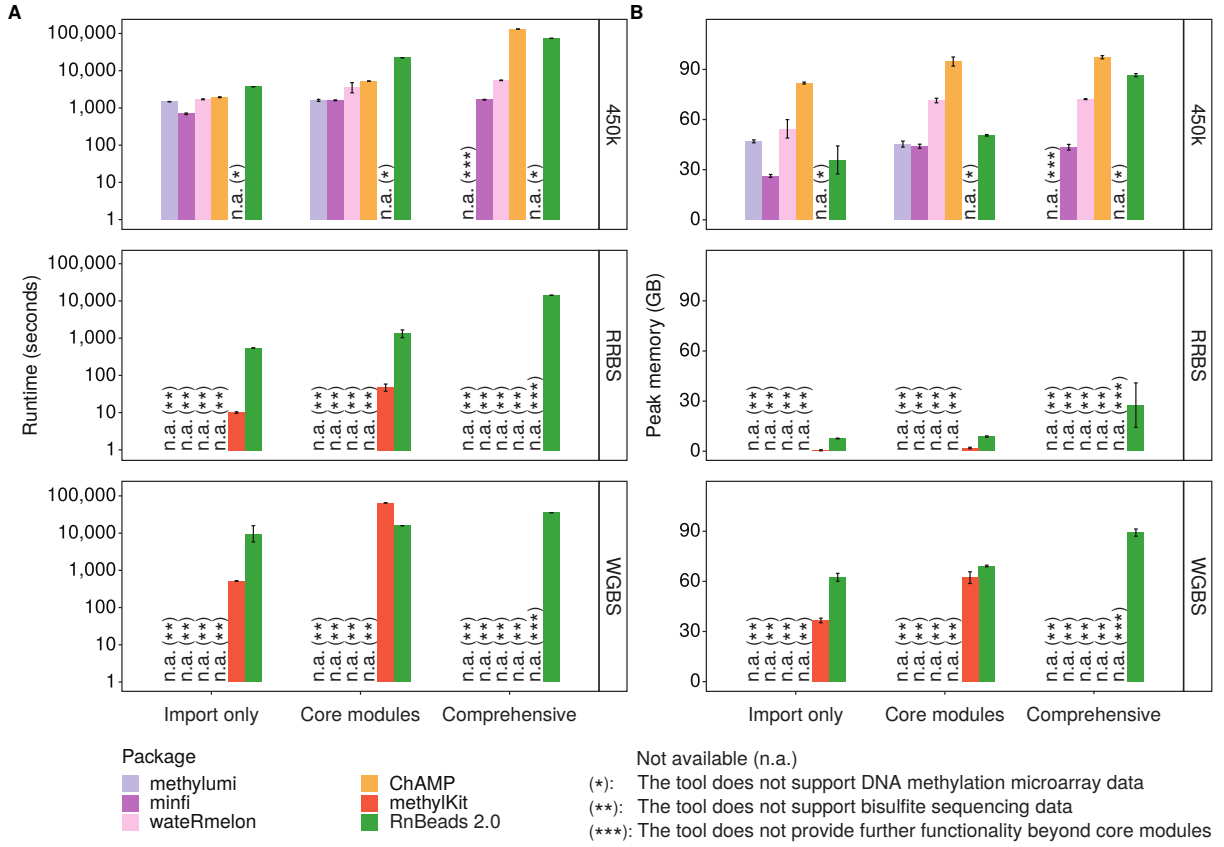


Figure 3.3: Comparing (A) runtime performance and (B) peak memory consumption of *RnBeads* with other tools for analyzing DNA methylation data including *methylumi*, *minfi*, *waterMelon*, *ChAMP*, and *methylKit*. Performance was evaluated on three datasets (450k, RRBS, WGBS), with three depths of analysis (import only, core modules, comprehensive analysis, see Table 3.1 for a more detailed description). The standard deviations were computed across three independent executions.

structure for downstream analysis. *ChAMP* and *RnBeads* provide a rich feature set that is comparable (Supplementary Table A.1), while *RnBeads* had reduced runtime and lower memory consumption in a setting with most features activated. *RnBeads* outperformed *methylKit* on the WGBS dataset in the core modules setting, while *methylKit* required less runtime and memory for the RRBS dataset. Since *RnBeads* stores larger matrices on disk rather than in main memory, re-formating the matrices for storage on disk consumes some runtime.

This brief benchmark showed that *RnBeads* has a runtime performance and memory consumption comparable to other available tools that provide a more limited feature set. In summary, the choice of software tool highly depends on the research questions asked, on the number of CpGs and samples analyzed, and should be carefully considered.

3.1.4 Discussion

We presented an extended version of the *RnBeads* software package with substantially extended modules and employed the new differential variability module to investigate tumor heterogeneity. Due to the new functionalities that we present here, *RnBeads* is one of the most com-

prehensive software suites available for performing DNA methylation analysis. *RnBeads* supports the analysis of microarray and bisulfite sequencing datasets and allows for the integration of datasets across technologies. However, in the current state of the package, the methylation signals are merely mapped to the same genomic position, without accounting for the specific properties of data generated using different technologies. These properties include different dynamic ranges of beta values/methylation values and the difference between bead and read coverage. More work is required to benchmark the integration across technologies. An open problem for the analysis of DNA methylation data is accounting for missing values in the data matrix. The updated version of *RnBeads* provides different imputation methods, but a comprehensive benchmark of these methods is missing.

Further extensions can be integrated into *RnBeads* as additional modules, such as *de-novo* identification of DMRs using methods including *BSmooth* [172], smoothing of DNA methylation values to better account for regional profiles [173], and support of single-cell bisulfite sequencing data. *RnBeads* can be used as an integral part of different epigenomic workflows due to its modular design and ease of use. In the context of this work, it will be used as a data processing tool, which stores DNA methylation data along with phenotypic information and genomic annotations. We envision that *RnBeads* will remain a widely-used software package due to continuous extensions, bug fixes, and updates.

In the use case that we present analyzing Ewing sarcoma samples, we found that analyzing either differential methylation or differential variability provides complementary information. For instance, we found increased epigenetic plasticity in cell lines in comparison to primary tumor samples using differential variability analysis. Notably, we used an aggregated level of DNA methylation computed across the sequencing reads for each of the CpGs and thus neglected read heterogeneity. We discuss how read heterogeneity can be used to further our understanding of within-sample heterogeneity on the Ewing sarcoma samples in Chapter 5.

3.2 DNA Methylation Dynamics During Aging

Aging is a process that affects virtually all organisms and investigating the human aging process is especially relevant. DNA methylation has recently emerged as a reliable biomarker for tracking the human aging process, since a subset of CpGs consistently loses or gains DNA methylation with increasing chronological age. This property of the methylome can be used to create reliable predictors for the chronological age of an individual [44, 46, 174]. Such predictors, often referred to as *epigenetic clocks*, use regularized linear regression such as elastic net regression optimized on a large training dataset to create age predictors based on a few hundred CpGs. For instance, the most popular and widely-used predictor of epigenetic age, the so-called *Horvath clock* is based on 353 CpGs. Recently, more advanced computational frameworks such as Map-Reduce were used to select CpGs predictive of the chronological age [175].

Predictors of the epigenetic age return estimates of the chronological age with a correlation to the true chronological age higher than 90% for healthy individuals [44, 46, 174]. The median absolute difference between predicted and annotated age is around three years. Since the methods have been trained on large datasets comprising healthy individuals, the estimate can be interpreted as the average chronological age of a person with the same DNA methylation pattern as the sample for which the age is to be estimated. The output is referred to as the epigenetic age. Epigenetic aging is accelerated (i.e., the epigenetic age is higher than the chronological age) in

different physiological and pathological states, such as HIV-1 infection [148, 176, 142] and predicts all-cause mortality [177]. Thus, these *clocks* are used as a surrogate for the effect of different environmental influences on the epigenetic age and for assessing the overall health state of the individual. Similar epigenetic age predictors have been developed for mouse [178, 179] based on bisulfite sequencing data. Epigenetic aging of cultured cells can be reverted by inducing pluripotency as to generate induced pluripotent stem cells (iPSCs), which have an epigenetic age of 0 [180]. Thus, algorithms predicting the epigenetic age can be used to test rejuvenation therapies or the effect of environmental influences on organism aging. Additionally, prediction of the epigenetic age has applications in forensic science [174], but can also be (ab)used for age determination of individuals with unknown chronological age.

3.2.1 Estimating DNA Methylation Age in *RnBeads*

The most widely-used method predicting the epigenetic age has been created by Steve Horvath and is often referred to as the *Horvath clock* [44]. To create this epigenetic age predictor, the author collected publicly available datasets obtained from healthy individuals that have been generated using the 27k and the 450k BeadChip with available age annotations. Using elastic net regression (cf. Section 2.6 [136]), 353 CpGs predictive of chronological age have been selected from the intersection of the CpGs available on the 27k and 450k array (around 21,000). The *Horvath clock* has gained popularity, since differences between the chronological age and the estimations were associated with various physiological and pathological states, including HIV1-infection [148], Werner syndrome [181], and physical and cognitive fitness [182]. However, the current state-of-the-art microarray is the EPIC BeadChip, and an evaluation of the epigenetic clock on the EPIC array showed consistent underestimation of the chronological age. This difference is likely caused by platform-dependent biases [183]. Notably, the current number of EPIC datasets that are publicly available is not sufficient to retrain the epigenetic clock.

Furthermore, no reliable epigenetic clock has been reported for bisulfite sequencing data. Thus, *RnBeads* employs platform-aware prediction of the epigenetic age. Datasets have been collected independently for the 27k and 450k bead array, and for RRBS. Data obtained from healthy individuals generated by the EPIC array are still scarce in open-access data hubs, and currently no age predictor is available for the EPIC array within *RnBeads*. Different pre-defined age predictors have been created in dependence to the platform to account for platform-specific biases and are readily available in the *RnBeads* package. The predictors are thoroughly described on the website¹³. More information can be found elsewhere [141]. *RnBeads* offers the functionality for training a new epigenetic age predictor on the EPIC array, as soon as a sufficient number of samples (more than 1,000) are available. In accordance with the observations in Dhingra et al. [183], we found consistent underestimation of the chronological age using *RnBeads'* 450k predictor on EPIC samples (Figure 3.4A,B).

In general, epigenetic age prediction in *RnBeads* is tissue- and cell-type-independent, given that the target tissue was also present in the training dataset. However, some tissues show altered epigenetic aging signals, which cannot be captured by standard age prediction models (Figure 3.4B). Samples obtained from the human cerebellum are particularly affected. These samples showed a substantially lower epigenetic than chronological age [184]. Since no cere-

¹³<https://rnbeads.org/ageprediction.html>

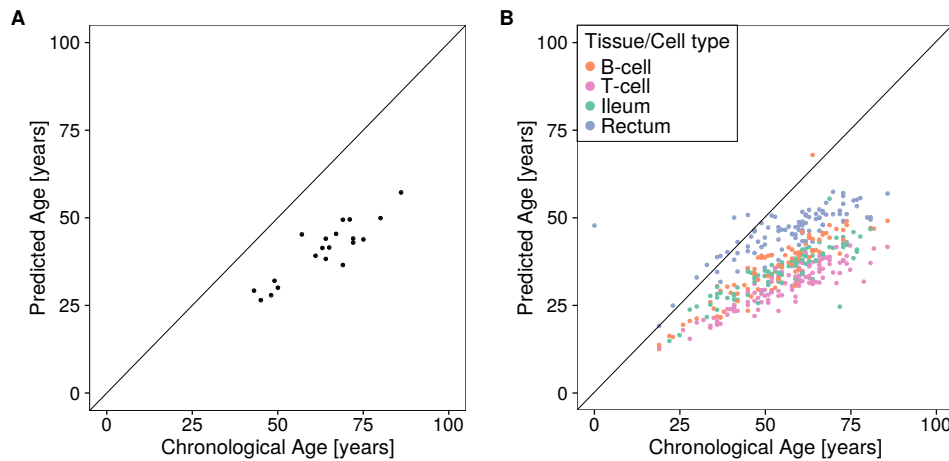


Figure 3.4: Scatterplot describing the predicted, epigenetic age (y-axis) and the annotated, chronological age (x-axis) for 21 whole blood samples from individuals in the SYSCID UCAM cohort (**A**, unpublished data) and for 409 samples from four different tissues and cell types in the CEDAR cohort (**B**, see Section 3.3.2). Epigenetic age prediction was conducted using the predefined predictor trained on 450k data.

bellum samples were present in the training dataset for the epigenetic age prediction module created in *RnBeads*, the module was unable to capture the distinct epigenetic aging patterns of cerebellum samples. Thus, this dataset constitutes a premier candidate for unsupervised domain adaptation; a statistical learning approach in which the target (test set) and source (training set) data do not follow the same distribution. Using a novel statistical framework employing unsupervised domain adaptation (called *wenda* for **w**eighte**d** **e**lastic **n**et for unsupervised **d**omain **a**daptation), we were able to accurately predict epigenetic age of samples obtained from human cerebellum [185].

3.2.2 DNA Methylation and Aging in Mouse

Similar to human samples, a plethora of murine datasets are publicly available. Additionally, lab mice are not affected by environmental influences similar to humans and the genetic background is more homogeneous, which removes two potential confounding factors from epigenomic studies. However, a microarray-based platform has only recently become available and currently no public datasets are available. Additionally, bisulfite sequencing of mice is only conducted in a limited number of laboratories worldwide. Most murine samples available in public databases have been assayed using RRBS, which generates an additional layer of uncertainty, since not all CpGs are well-covered in all samples due to the digestion with restriction enzymes. Thus, by increasing the number of samples the number of joint CpG sites decreases.

Using different mouse datasets, Stubbs et al. [178] generated an epigenetic age predictor for mice based on 329 CpGs. An important caveat of the study is the limited age range of the mice (0-41 weeks). Thus, we generated datasets for old mice (90 weeks) and applied the epigenetic age predictor proposed by Stubbs et al. [178]. The epigenetic age was estimated similar to those of ten-week-old mice (Supplementary Figure A.1) indicating that the predictor is not applicable to DNA methylation data generated on old mice (unpublished results together with the Institute of Pharmaceutical Biology at Saarland University).

3.2.3 Discussion

To investigate the aging process in both human and mouse, age prediction tools using DNA methylation data are of high relevance. Due to the association between accelerated epigenetic aging measured by an *epigenetic clock* and donor health states, these methods will remain an active research target. We provide a technology-aware age prediction method within *RnBeads*, which is yet to be extended for age prediction on EPIC array data after a large amount of healthy samples will have become available. Additionally, the predicted, epigenetic age should be treated with caution, since DNA methylation data can be affected by technical and biological variations. An important factor influencing epigenetic age estimates are cell-type-specific DNA methylation patterns. To alleviate cell-type-specific predictions, *wenda* can be used for predicting epigenetic age of tissues not present in the original training dataset. A reliable murine epigenetic age prediction tool is required and may be developed after a sufficient number (more than 1,000) of samples generated using the murine microarray will become available.

3.3 Identification of Tissue-Specific and Common Methylation Quantitative Trait Loci in Healthy Individuals Using *MAGAR*

3.3.1 Relationship Between Genotypes and DNA Methylation in MethQTL

As discussed in the previous section, DNA methylation can be affected by aging. Additionally, it can be influenced by sex, a range of environmental exposures [186, 187], and diseases including type I diabetes [74] and schizophrenia [69]. As another important factor, donor genotype has a strong influence on the global DNA methylation state, especially when a genetic alteration, such as a SNP, occurs at a CpG site. Using bisulfite treatment, unmethylated cytosines are converted into uracils (and further to thymines in subsequent PCR). Thus, one cannot differentiate between a genetic substitution of a cytosine base by a thymine and the bisulfite-induced sequence alteration. Without accounting for this, genetic alterations can be misinterpreted as DNA methylation differences. As a consequence, genomic regions containing annotated and predicted SNPs are typically removed from DNA methylation data (cf. Chapter 4).

In addition to genetic alterations affecting the CpG site itself, distant genetic variants can correlate with the DNA methylation state of a CpG. Such variants are referred to as methylation quantitative trait loci (methQTLs). These associations can range in distance from a few bases to several megabases, and also long-range interactions between different chromosomes have been reported [57, 188]. The definition of proximal methQTLs varies from 500 kb to 2 megabases (mb) distance between the CpG and the SNP [57, 188, 189]. MethQTLs co-localize with genetic variants associated with diseases and donor phenotypes (GWAS hits) including obstructive pulmonary disease [189], prostate cancer risk [190], osteoarthritis [191], immune-mediated disease [192], asthma [193], and smoking [187]. Furthermore, combining methQTLs with expression QTLs (eQTLs) enables the investigation of associations between DNA methylation and gene expression changes [59, 60, 54].

However, it remains elusive whether these methQTLs correlate with the DNA methylation level in a tissue- or cell-type-specific manner or whether they are largely tissue-independent. An earlier study used cultured cells including fibroblasts, T-cells, and lymphoblastoid cell lines to determine largely tissue-independent methQTLs. In contrast, the authors reported that

the association with gene expression changes was rather cell-type-specific [194], which is in line with recently identified cell-type-specific eQTLs [195]. In contrast, other studies reported largely cell-type-independent eQTLs [196]. Notably, methQTLs are typically determined using statistical models and tools that have been developed for eQTL analysis (e.g., *Matrix-eQTL* [197], *fastQTL* [198], or *GEM* [199]) without further accounting for the high correlation of DNA methylation states of neighboring CpGs.

To address this problem, we developed “Methylation-Aware Genotype Association in R” (*MAGAR*) a novel computational pipeline that performs methQTL analysis, while accounting for the properties of DNA methylation data. *MAGAR* defines clusters of neighboring CpGs according to their shared behavior (i.e., correlation) across samples to represent DNA methylation haplotypes and performs methQTL analysis for each of the correlation blocks independently. *MAGAR* has been implemented as an R-package and integrates with existing tools such as *fastQTL* [198], *RnBeads* (cf. Section 3.1), and *PLINK* [200]. Using *MAGAR*, we investigated sorted blood cell types (T-cells, B-cells) and composite bowel tissues (ileum, rectum) of healthy individuals. The identified methQTLs were also identified using additional samples and data from two published methQTL studies. We revealed both tissue-specific and common methQTLs using colocalization analysis. We identified more common than tissue-specific methQTLs and found that tissue-specific methQTLs were preferentially located in enhancer elements.

3.3.2 *MAGAR* - Methylation-Aware Genotype Association in R

MAGAR Package Overview

We developed *MAGAR* as a new computational framework to determine methQTLs from DNA methylation and genotyping data. *MAGAR* supports both sequencing-based assays including whole-genome (bisulfite) sequencing and microarray-based data. It is the first computational framework for performing methQTL analysis starting from raw DNA methylation and genotyping microarray data. The pipeline implemented within *MAGAR* comprises the following phases (Figure 3.5):

1. Data import and processing using established software packages such as *PLINK* [200], *RnBeads*, and *CRLMM* [201, 202]. Additional modules for quality control and standard processing using these packages are available to the user. *MAGAR* supports automated genotype imputation using the Michigan Imputation Server [124].
2. MethQTL calling for computing associations between genotype and a DNA methylation state is realized by a two-stage approach:
 - i. We define *CpG correlation blocks* as groups of CpGs that have similar (i.e., highly correlated) DNA methylation patterns across the samples to mimic DNA methylation haplotypes
 - ii. From each of these correlation blocks, a *tag-CpG* is selected as a representative of the block. Then, associations across the samples are computed between the DNA methylation states of the tag-CpG and the genotypes of all SNPs at a given distance using either a linear modeling strategy or with external software tools (e.g., *fastQTL* [198]). The output comprises SNP-CpG pairs that result in a p-value of the linear model (or the p-value returned by the external software tool) below a user-defined threshold.

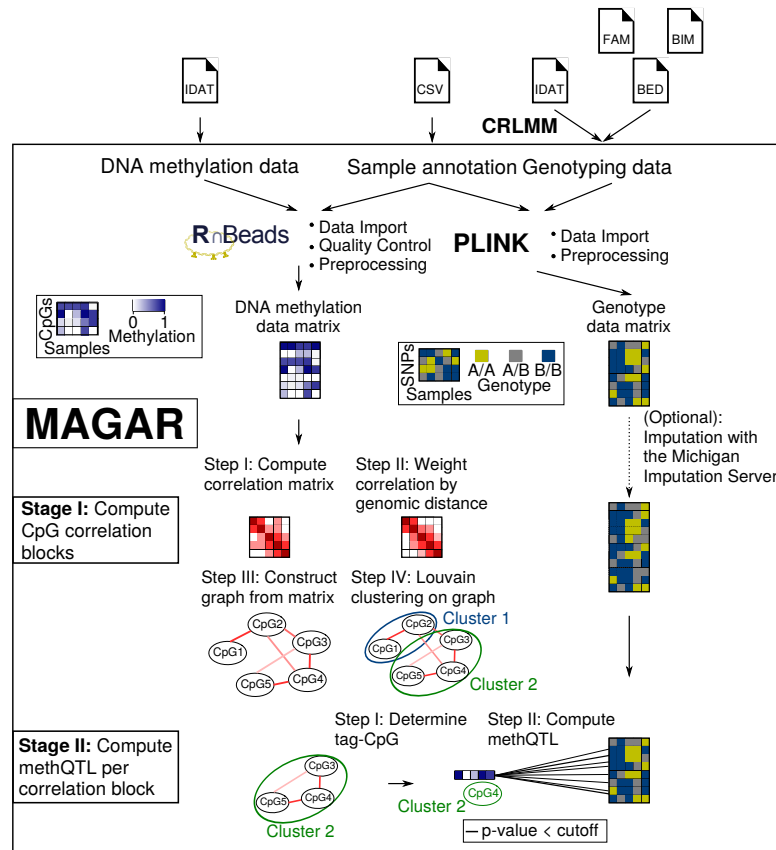


Figure 3.5: Overview of *MAGAR*. *MAGAR* is an R-package utilizing a two-stage protocol. After data import via external software packages, CpGs are clustered into CpG correlation blocks in a four-step procedure. In the second stage, methQTLs are called for each correlation block separately.

Data Import and Processing

DNA Methylation Data For DNA methylation data, we use *RnBeads* for data handling and processing. Microarray data is checked for data quality using *RnBeads*' reporting functionality. Further processing steps, such as CpG and sample filtering (e.g., removal of SNPs and cross-reactive sites) and data normalization, can be performed within *RnBeads*. Although we recommend *RnBeads* for data handling, *MAGAR* supports the output of alternative DNA methylation data processing tools if they provide single-CpG methylation calls.

Genotyping Data *MAGAR* supports microarray and sequencing data as input. Sequencing data has to be processed using genotyping pipelines [121] and converted into a format that is readable using *PLINK* (e.g., variant call format (VCF) files). For microarray data, *MAGAR* supports raw IDAT files as input and computes genotype calls through the *CRLMM* R-package [201, 202] (see also Section 2.5.1). As an optional step, genotyping data can be imputed using the Michigan Imputation Server [124]. Additional data processing such as filtering SNPs with many missing values or filtering according to the Hardy-Weinberg principle are conducted through *PLINK*.

MethQTL Calling

We define a methQTL interaction as a significant correlation between the SNP genotypes and the DNA methylation states of a CpG. MethQTL calling within *MAGAR* follows a two-stage workflow (Figure 3.5):

1. CpGs that have similar DNA methylation patterns across the samples are grouped together to form CpG correlation blocks.
2. A tag-CpG per correlation block is associated with all SNPs in a given genomic distance to compute methQTL interactions.

CpG Correlation Block Calling To compute CpG correlation blocks, i.e., CpGs that exhibit highly correlated DNA methylation patterns across the samples, we developed a four-step framework:

1. Compute the (Pearson) correlation coefficients between any pair of CpGs across the samples using the *bigstatsR* R-package [203] for each chromosome separately and use the resulting correlation matrix as the similarity matrix. Similarities for CpGs with correlations lower than 0.2 (package parameter: `cluster.cor.threshold`) are set to zero. Since matrices can grow too large to fit into main memory of standard machines, the CpGs are split per chromosome into equally-sized smaller groups until a maximum number of CpGs (i.e., rows of the data matrix) per computation is achieved (here 40,000 CpGs, parameter: `max.cpgs`).
2. Since also distant CpGs can exhibit high correlation (e.g., those present in CpG islands), we penalize the similarity of two CpGs according to their genomic distance. Thus, we weight the genomic distance between any CpG and its genomic neighbors according to a Gaussian centered at the CpG of interest with standard deviation 3,000 bp (parameter: `standard.deviation.gauss`). Additionally, the similarity between any pair of CpGs further apart than 500 kb is set to zero (parameter: `absolute.distance.cutoff`). Optionally, functional annotations such as those from the Ensembl Regulatory Build [8] or DNA methylation-based segmentation [153] can be used to weight the similarities.
3. Construct the associated weighted graph from the similarity matrix, where the weights of the edges correspond to the similarities between the two CpGs.
4. Employ Louvain clustering (see Section 2.6.5 [139]) using the *igraph* R-package [204] to the weighted graph to obtain clusters of CpGs that are highly correlated. The obtained clusters are defined as the CpG correlation blocks.

The parameters presented here are available as package options to the user. The default parameters have been evaluated using simulations for EPIC and 450k data (see Section 3.3.3).

Associating SNPs with CpG Correlation Blocks To determine whether the DNA methylation state of a CpG correlation block is correlated with a SNP genotype, we first determine a tag-CpG per correlation block as the medoid of all CpGs in the correlation block. Alternative tag-CpG selection methods (e.g., the average DNA methylation profile across the CpGs) are available

through the package parameter `representative.cpg.computation`, but we used the `medoid` here due to the better interpretability in comparison to the other methods. In the next step, all SNPs closer than 500 kb to the tag-CpG are considered and a univariate least-squares regression (`lm` R function) model is calculated for each pair of SNPs/CpGs individually using the genotypes (encoded as 0=homozygote reference/major allele, 1=heterozygote, 2=homozygote alternative/minor allele) as the features and the CpG methylation state as the output. Further covariates can be included in the linear model. We decided to use a univariate linear regression model for improved interpretability in comparison to a multivariate (regularized) linear regression using all eligible SNPs as the input. Alternatively, *fastQTL* [198] can be used to compute associations between tag-CpGs and SNPs. The obtained p-values, slopes (effect sizes, betas), and standard errors of the linear model are used for further analysis.

Package Options and Modularity

MAGAR is a modular software package that allows for easy integration with additional tools. Different flavors of the analysis can be specified using the package's rich option set (Table 3.3). For instance, CpG correlation blocks depend on various parameters including the correlation threshold between two CpGs or the standard deviation of the Gaussian distribution. The option setting can be tailored to the dataset at hand. For instance, CpG correlation block calling can be deactivated, resulting in the analysis scheme implemented by most published *methQTL* studies, i.e., associating each CpG with a SNP individually. Additionally, *MAGAR* allows for setting the parameters of the different software tools that are used for data processing (e.g., *RnBeads*, *PLINK*). To facilitate analysis of large-scale datasets, *MAGAR* supports multi-core processing and automatic distribution of jobs across the nodes of an HPC cluster (SGE and SLURM architecture supported). *MAGAR* is available from GitHub¹⁴.

Data Simulation for Determining *MAGAR*'s Default Parameters

MAGAR is a modular software package that allows multiple parameters to be set for the dataset at hand. We simulated data to determine reasonable default parameters for the two stages of the package independently.

Correlation Blocks As the first part of the *MAGAR* package, CpGs are grouped together according to their correlation of DNA methylation values across the samples. The process of defining correlation blocks depends on three parameters: the correlation threshold, the standard deviation of the Gaussian distribution, and the absolute distance cutoff (Table 3.3). To determine reasonable default values for the parameters of the correlation block calling of the package and to validate that the CpG clustering step is reasonable, we simulated DNA methylation data. More specifically, we simulated methylation data and artificially introduced clusters of highly correlated CpGs into the data. Then, we assessed whether the CpG correlation blocks returned by *MAGAR* reflect the simulated clusters of correlated CpGs. By using different values for the parameters, we were able to assess the parameter setting that best reflects the correlation of CpGs in the simulated data (see Section 3.3.3).

¹⁴<https://github.com/MPIIComputationalEpigenetics/MAGAR>

Table 3.3: Overview of *MAGAR*'s option setting

Option	Function	Default	Selection of Default
<code>cluster.cor.threshold</code>	Threshold below which the similarity in the similarity matrix to compute the CpG correlation blocks is set to zero	0.2	Simulations
<code>standard.deviation.gauss</code>	Standard deviation (in bp) of the Gaussian distribution used to penalize similarities of CpGs according to their genomic distance	3,000	Simulations
<code>absolute.distance.cutoff</code>	Maximal distance (in bp) between two CpGs, similarities of pairs of CpGs with higher distance are set to zero	500,000	Simulations
<code>representative.cpg.computation</code>	Method for selecting a tag-CpG from all CpGs present in a correlation block	median	Interpretability
<code>max.cpgs</code>	Maximum number of CpGs used to construct the similarity matrix	40,000	Computational Feasibility
<code>correlation.type</code>	Method for computing the correlation between two CpGs	pearson	Computational Feasibility

To simulate methylation data, 1,000 neighboring CpGs were randomly selected from a uniform distribution out of all the CpGs present on the Illumina EPIC array. 1,000 CpGs were selected as a compromise between selecting all CpGs and computational feasibility of executing multiple simulations. We explored different settings for the parameters available in *MAGAR*. First, we explored the influence of the correlation threshold parameter (values tested: 0 to 1, 0.05 steps), which specifies the level of correlation between two CpGs that results in a similarity of zero in the similarity matrix. Second, the standard deviation of the Gaussian distribution (values explored: 2,000 bp to 4,000 bp, 100 bp steps) specifies the width of the Gaussian distribution that penalizes similarities of distant CpGs. The selection of the values for the distance was based on the distances of CpGs on the microarray and should reflect high-to-low penalization of the genomic distance. Last, the absolute distance cutoff (100 kb to 1 mb, 100 kb steps) sets similarities for long distances between the CpGs to zero. This value was selected, since it also reflects the distance between the SNP and the CpG that we selected (500 kb).

The parameters were tested using 100 simulated datasets per parameter setting. For each of the simulated datasets, we explored the three parameters sequentially and fixed the remaining parameters to the values estimated in the other simulations (starting with 3,000 bp standard

deviation and absolute distance cutoff 500 kb for selecting the correlation threshold). Methylation data was simulated by repeatedly drawing from a beta-binomial distribution with success probability 0.4 and over-dispersion parameter 0.1 in order to reflect the typical bimodal distribution of DNA methylation data. For each simulated dataset, we selected the number of clusters randomly between 200 and 500, while choosing the cluster size individually for each cluster between one and ten CpGs. These values should reflect clusters of many CpGs (such as those expected at CpG islands), but also clusters with only few CpGs and were motivated from the distribution of CpGs on the EPIC array. The clusters had identical methylation patterns across the CpGs in the cluster and across the samples. We introduced a Gaussian error for each CpG individually (standard deviation 0.05) to introduce noise into the clusters, which is motivated from our experience on DNA methylation data and the technical noise found in data generated using the EPIC array.

To assess *MAGAR*'s performance, we executed its first stage and compared the number of expected clusters (i.e., the randomly selected number of clusters) with the number of clusters returned by *MAGAR* (see Section 3.3.3). For estimating the parameters for 450k data, we exclusively used CpGs present on the 450k array and for bisulfite sequencing data we used all CpGs available in the human genome reference version 'hg19'.

Validating MethQTL Calling To validate the methQTL calling stage of *MAGAR*, we first generated methylation data as described above. Next, we randomly selected 2,000 SNPs that are located more closely than 500 kb from the CpGs selected. We selected more SNPs than CpGs, since microarray-based technologies for genotyping cover more SNPs than DNA methylation microarrays cover CpGs. For those SNPs, we drew the minor allele frequency from a negative binomial distribution (parameter success probability: 0.4) and set the alleles accordingly. We selected the negative binomial distribution, since it most-closely reflected the distribution of genotypes in our experimental data. SNP genotypes (α) were encoded as 0=homozygote reference allele, 1=heterozygote, and 2=homozygote alternative allele similar to the standard encoding of *MAGAR*. For each of the 100 simulated experiments that we conducted, we introduced 100 interactions between the genotype of a SNP and the DNA methylation state of a CpG into the data using a randomly selected effect size τ (drawn from a normal distribution with mean 0.2 and standard deviation 0.05). We decided to include only a small number of methQTLs in order to have only few interactions between SNPs and CpGs as we would expect in real data. The sign of the effect size τ was randomly selected as positive or negative, respectively. Similar to the simulation above, we introduced a Gaussian error ϵ into the DNA methylation data. The DNA methylation state β was modified according to:

$$\beta_{\text{CpG}}^{\text{new}} = \beta_{\text{CpG}}^{\text{old}} + \alpha_{\text{SNP}} \times \tau + \epsilon$$

We then computed sensitivity and specificity for the CpGs and SNPs independently to assess whether the package successfully identified methQTLs (see Section 3.3.3).

Datasets

The datasets used throughout this project have been generated in the context of the SYSCID project¹⁵. The CEDAR (Correlated Expression Disease Association Research [53]) cohort dataset comprises 164 individuals, and we had microarray-based genotyping data available for 163 individuals as described earlier [53]. More specifically, healthy individuals were recruited at the University Hospital in Liège and bowel biopsies as well as blood draws were obtained. The biopsies were obtained from rectum (RE) and ileum (IL), and blood cells were sorted into CD4-positive T-cells and CD19-positive B-cells. DNA methylation data was generated using the Illumina EPIC microarray by Gilles Gasparoni from the Genetic/Epigenetics department, and we used this dataset as the discovery cohort. In addition, we used additional samples from the CEDAR cohort as second dataset comprising additional 197 donors (16 overlapping with the earlier ones) with transverse colon biopsies (n=191) and CD14-positive monocytes (n=192) as a validation cohort. DNA methylation data for the validation cohort was generated using the Illumina 450k array.

MAGAR Analysis of the CEDAR Cohort

DNA Methylation Data We used *MAGAR*, which internally uses *RnBeads*, for processing raw IDAT files obtained on the CEDAR cohort samples. A subset of samples (13 B-cell samples, 1 T-cell sample) was removed from the discovery cohort, since the samples exhibited substantially lower technical quality. CpGs were filtered for SNPs annotated in dbSNP [113], for sites on the sex chromosomes, and for potentially cross-reactive sites [91]. Further quality-based filtering of CpGs was conducted using *RnBeads*' GreedyCut algorithm [112]. Data was normalized using the "dasen" method from the *wateRmelon* R-package [116]. As outcome of the filtering procedure, 659,464 CpGs were retained for the analysis. The immune cell infiltration was estimated using the LUMP algorithm [149] based on 44 CpGs that are particularly hypomethylated in immune cells, 34 of which are available in the CEDAR dataset. For the validation dataset, we used analogous processing options, removed one sample from the 383 samples due to lower technical quality, and retained 353,388 from the 485,777 CpGs available on the microarray.

Genotyping Data Genotyping microarray data was imported into *MAGAR* and genotypes were called using the *KRLMM* algorithm implemented in the *CRLMM* R-package [201, 202] with default parameters (cf. Section 2.5.1). Genotypes were imputed using the Michigan Imputation Server [124] and the following parameters: Reference panel: "hrc-r1.1", phasing method: "shapeit", population: "eur". Imputation was performed for all 163 unique donors simultaneously and the outcome of the procedure yielded 39,127,678 SNPs. Imputed data was exported to *PLINK* [200] for further processing. We filtered for SNPs with a Hardy-Weinberg equilibrium exact test p-value below 0.001, a maximum number of missing values across the samples of 10%, and with minor allele frequency below 5%. Additionally, we removed samples with more than 5% missing genotypes. As an outcome of the filtering procedure, no sample was removed and 5,436,098 SNPs were retained.

MethQTL Analysis We employed *MAGAR* on an HPC cluster to compute methQTLs for each of the tissues/cell types of the CEDAR cohort dataset independently (Figure 3.6). Notably, we

¹⁵<http://syscid.eu/>

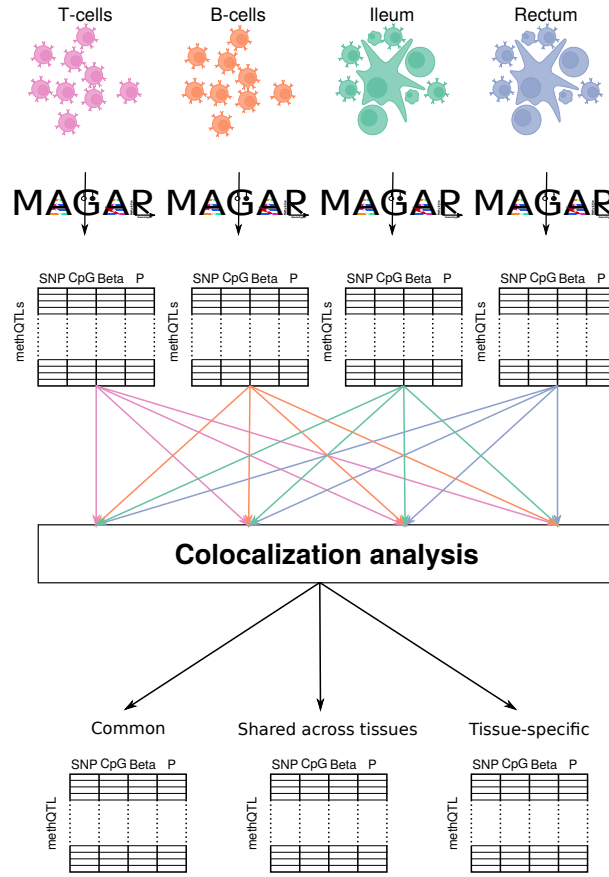


Figure 3.6: Identifying common and tissue-specific methQTLs through colocization analysis. To define tissue-specificity, we employed *MAGAR* on the four tissues/cell types independently. The methQTL statistics were combined across the four tissues in pairwise colocization analysis to define common and tissue-specific methQTLs, as well as methQTLs shared across several tissues.

used sex, age, body mass index (BMI), smoking habit, alcohol intake, ethnicity, and the first two principal components computed on the genotype data as covariates in the analysis. *MAGAR* returns a table of methQTL summary statistics (p-values, slopes), which can be further filtered according to a user-defined p-value cutoff. Throughout this analysis, we termed methQTLs significant, if they passed a genome-wide Bonferroni-adjusted p-value cutoff of 8.65×10^{-11} in the summary statistics returned by *MAGAR*. We computed the p-value cutoff as follows: We identified 82,271, 69,219, 75,779, and 76,109 correlation blocks for T-cells, B-cells, ileum and rectum samples, respectively (Supplementary Figure A.3). On average, each CpG was tested for association with 1,905 SNPs, which results in a Bonferroni-adjusted p-value cutoff of:

$$\frac{0.05}{(82,271 + 69,219 + 75,779 + 76,109) \times 1905} = 8.65 \times 10^{-11} \quad (3.1)$$

For each CpG that was affected by more than one methQTL, we selected the SNP with the lowest p-value as the lead-SNP and discarded the interactions with other SNPs.

Defining Tissue-Specific MethQTLs

To determine whether the effects observed in the four tissues independently were shared across the samples, we employed colocalization analysis. More specifically, we used Summary-data-based Mendelian Randomization (SMR) and Heterogeneity in Dependent Instruments (HEIDI) analysis [205] implemented in the GWAS-MAP¹⁶ software tool. Briefly, SMR is a statistical test that indicates whether two traits (here CpG methylation states in two tissues) are significantly associated with the same genetic locus. The test is an extension of Mendelian Randomization (MR), which is used to test for a causal relationship between two traits using an instrumental variable. While classical MR requires that the two traits are measured on the same samples, these can be investigated in distinct samples or studies using SMR. The input to the SMR test are methQTL statistics (i.e., p-values, slopes of the regression) obtained in two scenarios, and it returns a test statistic that indicates whether the effect observed in the two scenarios is associated with the same genetic locus. SMR analysis determines whether the same genetic effect leads to the methQTL results that we obtained in the two tissues, but cannot discern pleiotropy from linkage (cf. Supplementary Figure A.2). Thus, for the SNPs that pass the SMR test, we employed the HEIDI test in a second step to test whether the observed effects are likely driven by pleiotropy. Briefly, the HEIDI test utilizes linkage (correlation) information of SNPs from a reference panel (e.g., the 1,000 genomes project [121]) to determine whether the observed heterogeneity in the methQTL statistics are more likely caused by linkage than by pleiotropy. By using colocalization analysis through SMR and HEIDI, we were able to determine whether the methQTLs identified in the four tissues/cell types independently were shared or tissue-specific. We employed colocalization analysis for all pairs of tissues/cell types to determine shared methQTLs (Figure 3.6).

We selected those CpGs for colocalization analysis, which were selected as tag-CpGs in at least two tissues and that had a significant association with a lead-SNP (p-value smaller than 8.65×10^{-11}) at least in one tissue. Then, anchoring the analysis in the tissue showing the significant association, we performed the SMR test to detect if the same lead-SNP may be associated with the same CpG in any of the other tissues. In case the same lead-SNP was identified in more than one tissue, the tissue/cell type with the lowest p-value was used as the starting point of the SMR analysis. In total, we performed 4,253 tests. The SMR p-values were adjusted for multiple testing using the Benjamini-Hochberg [206] method and we used a p-value cutoff of 0.05. In case the methQTLs measured in two tissues are significant according to the SMR test, this is an indication that the CpG methylation states are correlated with the same SNP in the two tissues. Thus, we used the p-value of the SMR test as an indication for the shared effect of methQTLs in the two tissues.

For CpGs that passed the SMR test, we applied the HEIDI test to discern pleiotropy, i.e., that the SNP correlates with two traits independently, from linkage, i.e., that there are two independent, but correlated SNPs each of which is associated with one of the traits (cf. Supplementary Figure A.2). We defined all those pairs of methQTLs with a p-value of the HEIDI test higher than 0.05 as pleiotropic interactions. The methQTLs that passed the SMR p-value cutoff and failed the HEIDI test were defined as shared across the two tissues. The methQTLs shared across *all* pairwise comparisons according to the colocalization analysis were termed *shared methQTLs*. Additionally, those shared methQTLs with a p-value below 8.65×10^{-11} in

¹⁶<https://polyknomics.com/solutions/gwas-map-biomarker-and-intervention-target-discovery-platform>

the methQTL analysis for all tissues were termed *common methQTLs*.

The methQTLs that either fail the SMR test or that pass the SMR test, but also pass the HEIDI test were defined as tissue-specific methQTLs. Tissue-specificity was defined for each tissue individually. Finally, three classes of methQTLs were defined: tissue-specific, shared, and common methQTLs. Colocalization analysis was performed using GWAS-MAP by Yurii Aulchenko and Tatiana Shahkova.

Characterizing Tissue-Specific and Tissue-Independent MethQTLs We used all methQTLs present in any of the four tissues and compared the effect sizes (slope of the regression), p-values, and the distance between the CpG and SNP of the tissue-specific with the methQTLs shared across the tissues. The effect size reflects the change in the DNA methylation state of the CpG that occurs from the homozygote (reference allele) to the heterozygote (alternative allele) genotype for the SNP. Additionally, we selected different functional annotations of the genome, such as Ensembl genes (version 75), associated promoter regions (defined as 1.5 kb upstream and 0.5 kb downstream of the TSS), and different functional categories according to the Ensembl regulatory build [8]. Then, we overlapped the shared/tissue-specific methQTLs with those annotations using the *GenomicRanges* [207] R-package and computed odds ratios and (one-sided) Fisher-exact test P values to investigate enrichment with respect to the functional annotations in comparison to all identified methQTLs as the background. Last, we used the LOLA tool [152] to compute enrichments regarding various additional functional annotations from databases such as Cistrome [170], CODEX [169], or ENCODE [168]. Here, we used all CpGs/SNPs that were eligible for methQTL analysis as the background for the enrichment.

Validation of MethQTLs

For further validation of the methQTLs identified above, we used 191 transverse colon samples and 192 monocyte samples from the CEDAR cohort assayed using the Infinium 450k microarray. Genotyping and DNA methylation data was processed analogously to the discovery cohort and methQTLs were called at the p-value cutoff 9.84×10^{-6} . We aimed to validate the 2,508, 696, 1,010, and 868 methQTLs that we identified in the four tissues/cell types and thus computed the p-value cutoff as:

$$\frac{0.05}{2508 + 696 + 1010 + 868} = 9.84 \times 10^{-6}$$

We used sex, age, BMI, smoking habit, alcohol intake, ethnicity, and the first two principal components computed on the genotype data as covariates. The resulting methQTLs were compared with the shared and tissue-specific methQTLs detected in the discovery cohort, respectively. Additionally, we obtained methQTL data in tabular form from two studies identifying methQTLs in peripheral blood [57] and fetal brain samples [208], respectively. The two studies identified 52,918 (blood) and 16,811 (fetal brain) methQTLs. We only used unique SNPs with a p-value lower than 8.65×10^{-11} to match our criteria. To determine whether the detected overlap was larger than expected by chance, we used Fisher's exact test using all SNPs that have been used as input to the methQTL calling as the background set.

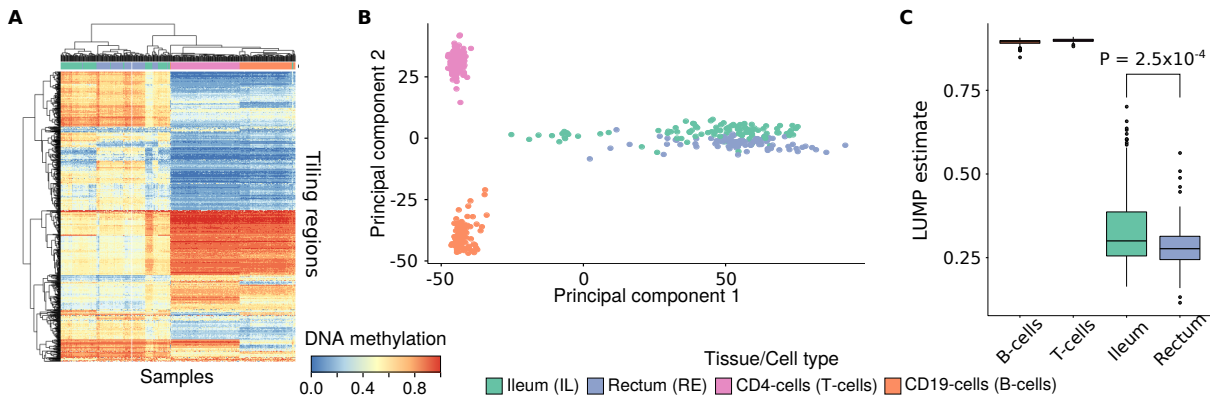


Figure 3.7: Cell-type specific DNA methylation patterns in the discovery data set. **A:** Heatmap (blue low, red high DNA methylation levels) of the 1,000 most variably methylated genome-wide tiling regions of size five kb. Hierarchical clustering of samples and tiles was performed using Euclidean distance and complete linkage. **B:** PCA plot of genome-wide DNA methylation data at the single CpG level. Shown are the first two principal components. **C:** Boxplots showing the LUMP estimates for the overall immune cell content of the different cell types/tissues. The p-value was computed using a two-sided t-test.

3.3.3 Distinct Biological Properties of Tissue-Specific and Common MethQTLs

Strong Cell-Type-Specific DNA Methylation Signatures in Bowel Biopsies and Purified Blood Cell Types

The data set that we used for the discovery of methQTLs comprised 409 samples from either ileum (IL, $n=98$) and rectum (RE, $n=95$) tissue biopsies, as well as the FACS-sorted blood cell types CD4-positive T-cells ($n=119$) and CD19-positive B-cells ($n=97$). DNA methylation data was available across all four tissues/cell types for 29 individuals. Average DNA methylation levels across all CpGs in genome-wide windows of size five kilobases revealed a strong cell-type-specific signal that discriminates the blood cell types from the biopsies. Overall, the tissue biopsies showed an enhanced variation in comparison to the purified blood cell types indicating that increased cell-type heterogeneity goes along with a higher variation of DNA methylation patterns (Figure 3.7). We estimated the overall immune cell content of a sample using the LUMP algorithm (Figure 3.7C) to better understand the origins of cellular heterogeneity within the biopsy samples. While LUMP estimates were uniformly close to one for the two blood cell types as expected, they substantially varied across the biopsy samples. In line with previous reports [209], significantly higher immune cell content was observed in ileal compared to rectal samples.

MAGAR: Genome-Wide Analysis of MethQTLs

Defining methQTLs is important to interpret genetic variants associated with diseases and can help to illuminate the association between genetic alterations and gene expression changes. Thus, we are interested in defining statistically significant associations as methQTL based on DNA methylation and genotyping data. To alleviate the methQTL identification process, we outline a new R-based framework, MAGAR (Methylation-Aware Genotype Association in R) that provides a comprehensive suite of tools that enable methQTL analysis in a manner that

is aware of the structure of DNA methylation data (Figure 3.5). Notably, *MAGAR* is the first package that performs data processing of raw (i.e., IDAT files) DNA methylation and genotyping data before returning data formatted for methQTL analysis.

MAGAR is a flexible software package that allows users to adapt the analysis to their dataset through different option settings. We determined *MAGAR*'s default parameters using a simulation strategy. More specifically, we estimated the default parameters for the correlation block calling (Figure 3.8A-C) and methQTL calling stage independently (Figure 3.8D). We selected the parameters such that the number of clusters (i.e., CpG correlation blocks) returned by *MAGAR* matches the number of clusters that have been simulated. The simulation experiments returned 0.2 as a reasonable parameter for the correlation threshold, which determines at which level of correlation between the two CpGs an edge is removed from the similarity graph. Additionally, 3,000 bp was selected as the standard deviation of the Gaussian distribution, which weights the similarity between two CpGs according to the genomic distance. Notably, higher values for the parameter would more closely reflect the number of simulated clusters in the data. However, we decided to fix the parameter at 3,000 bp, since we expect that generating fewer cluster (i.e., more singletons) will not have a negative influence on the identified methQTLs. Lastly, we found that the distance cutoff only mildly influenced the number of clusters generated, and we determined 500 kb as the distance at which a connection between two CpGs in the graph is removed. This value also matches the maximum distance between the SNP and the CpG that we selected. To validate whether methQTLs are reliably detected using our package, we artificially introduced interactions between SNPs and CpGs into our simulations. We found high sensitivity and specificity for methQTLs to be detected by *MAGAR* across the 100 simulated datasets that we generated (Figure 3.8D). Notably, *MAGAR* was designed to detect reliable methQTLs with few false positive results. Thus, the focus of *MAGAR* is on specificity rather than on sensitivity, but the user can tradeoff between sensitivity and specificity through the p-value cutoff.

Next, we employed *MAGAR* on the discovery cohort and investigated tissue-specificity of the identified CpG correlation blocks. To that end, we employed *MAGAR* on the four tissues/cell types individually and compared the resulting correlation blocks. Although we found that most correlation blocks and their respective tag-CpGs were tissue-dependent, some of the correlation blocks, which we computed for each of the tissues independently, were shared across multiple cell types (Supplementary Figure A.3). Notably, the overlap between tag-CpGs across the different tissues/cell types was larger than between the correlation blocks indicating that only parts of the correlation blocks were distinct for the different data sets.

Using *MAGAR* we combined the ileal, rectal, T-cell, and B-cell methylation data with genotype data and calculated methQTL statistics for each cell type/tissue independently. To determine significant methQTLs, we selected a Bonferroni-corrected p-value cutoff of 8.65×10^{-11} . As a result, we found 696, 2,508, 1,010, and 868 methQTLs for CD19+ B-cells, CD4+ T-cells, ileal, and rectal biopsies, respectively (Figure 3.9). To validate the methQTLs, we used additional samples from monocytes and transverse colon from the CEDAR cohort. Additionally, we obtained published methQTL results from two studies (blood [57] and fetal brain [208]) and compared them with the identified methQTLs. Note that the validation cohort and the published studies used DNA methylation data generated on the 450k microarray, which comprises fewer CpG sites. Thus, we excluded those methQTLs from the comparison that associated with a CpG site that is exclusively present on the EPIC array. We partially identified the methQTLs

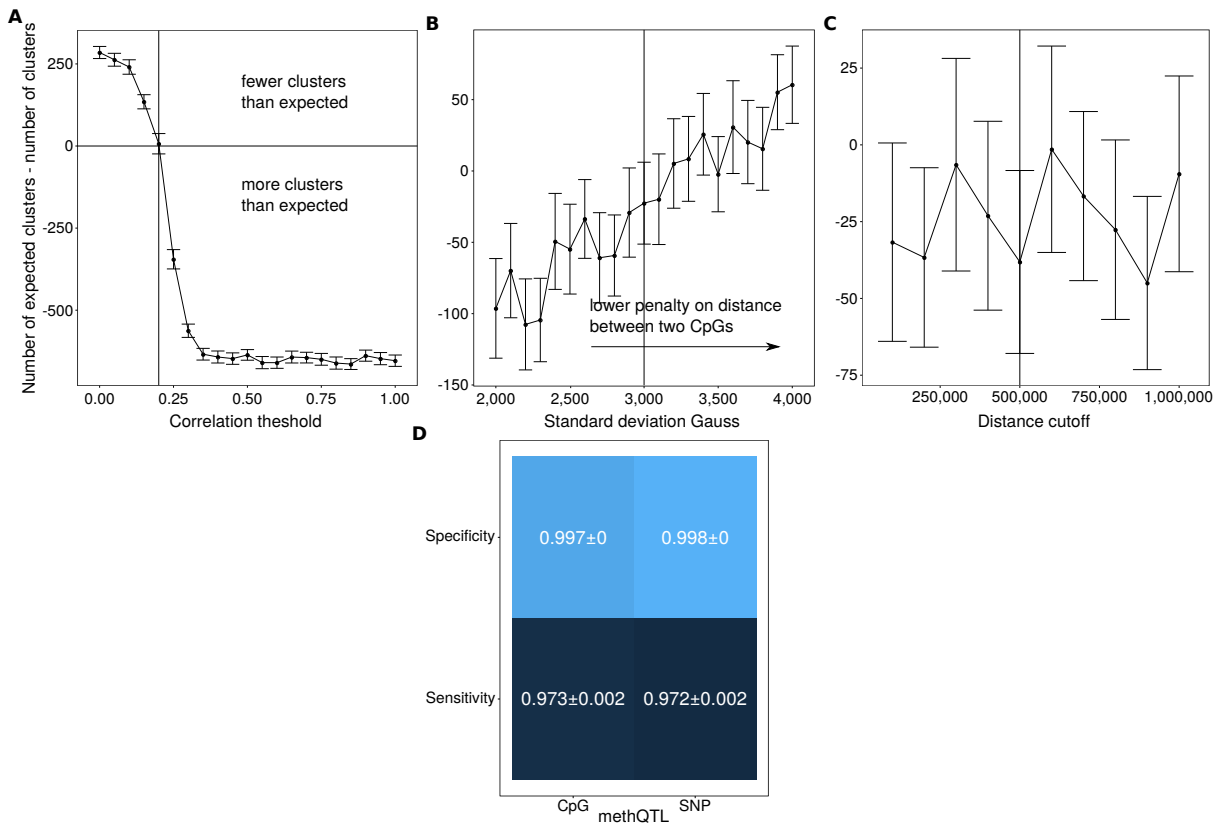


Figure 3.8: Validating *MAGAR* and its parameters using simulated data. **A:** Difference between the number of expected clusters and the number of clusters generated by the package in comparison to the correlation threshold parameter. Higher values on the y-axis indicate that the clusters/correlation blocks generated by the package are too large. The solid line indicates the selected default value of the parameter for EPIC data. The error bars indicate two times the standard error computed across the 100 simulated datasets per parameter setting. Effect of the standard deviation of the Gaussian distribution (**B**) and the absolute distance cutoff (**C**) on the number of clusters. **D:** Sensitivity and specificity of *MAGAR*'s methQTL calling in simulated data for CpGs and SNPs independently. Shown is the mean and the standard error across the 100 simulated datasets.

also in the validation cohort (Figure 3.9B) and in the published data (Figure 3.9C). As expected, the overlap of the methQTLs identified in B- and T-cells with the methQTLs identified using whole blood was higher than with those identified in fetal brain samples (Figure 3.9C).

Identification of Common MethQTLs Through Colocalization Analysis

MAGAR's output was further analyzed to discern tissue-specific from common methQTLs. More specifically, we applied colocalization analysis that uses summary statistics from two association studies (here methQTLs in two tissues) to determine if an association of two traits (here CpG methylation states) to the same genetic region is significant and is likely to be caused by the same pleiotropic genetic variant. Colocalization was examined using SMR analysis followed by the HEIDI test [205].

We only included methQTLs in the analysis that were significant at 8.65×10^{-11} in at least one tissue. The analysis is anchored at the tissue, where the methQTL showed a significant

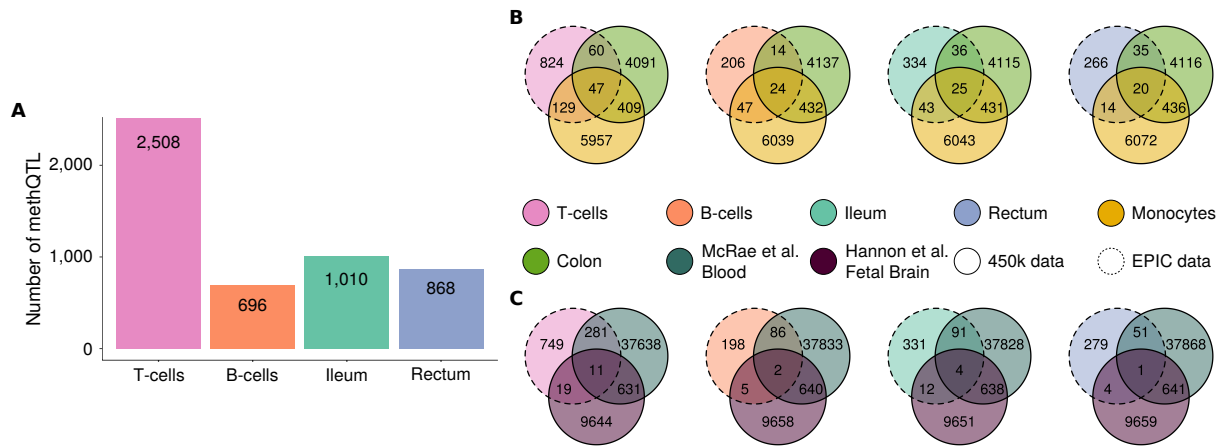


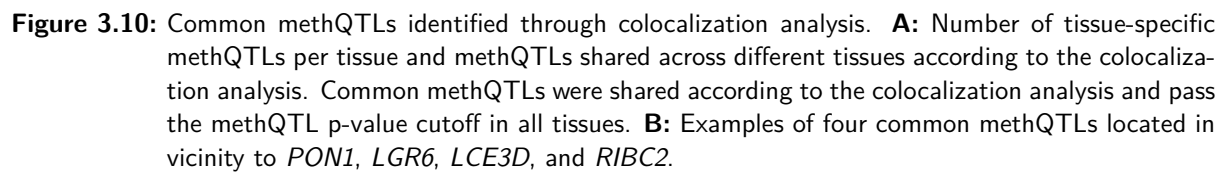
Figure 3.9: MethQTL results returned by *MAGAR*. **A:** Number of methQTLs identified by *MAGAR* for T-cells, B-cells, ileal, and rectal samples. Overlap between the methQTLs identified per tissue/cell type with methQTLs identified in the validation cohort (**B**) and in published methQTLs from blood [57] and fetal brain samples [208] (**C**). The methQTLs were reduced to those methQTLs affecting CpGs present on the 450k microarray.

association and the methQTL statistics were compared with those in the other tissues. We defined those methQTLs as shared between two tissues/cell types that pass the SMR-test at FDR-adjusted p-value cutoff 0.05 and have a HEIDI test nominal p-value larger than 0.05. These methQTLs are likely correlated with the same genetic variant and the shared association is likely caused by a single pleiotropic variant rather than two linked signals. Colocalization analysis was conducted for all pairs of cell types/tissues (Figure 3.6) and we define three classes of methQTLs:

1. Common methQTLs are shared across all the tissues/cell types according to the colocalization *and* have a methQTL p-value below 8.65×10^{-11} in all tissues
2. Shared methQTLs are shared across all the tissues/cell types according to the colocalization analysis
3. Tissue-specific methQTLs are only present in one of the tissues/cell types and not shared in any pairwise comparison

We found that 16 methQTLs were shared across all of the pairwise comparisons and have a methQTL p-value below the threshold and are thus common methQTLs (Figure 3.10A). The common methQTLs included well established methQTLs and eQTLs, such as the ones present in the *PON1* [210], *LGR6* [211], and *RIBC2* [212] loci (Figure 3.10B). We found substantially more methQTLs shared across different tissues than tissue-specific methQTLs. Most tissue-specific methQTLs were exclusively found in CD4 T-cells (Figure 3.10A), and similar numbers of tissue-specific methQTLs (78, 75) were identified for ileal and rectal biopsies, respectively. Due to the definition above, common methQTLs are a subset of the shared methQTLs.

We further investigated the identified common and shared methQTLs using the validation cohort. Notably, the validation cohort samples have been assayed using the 450k array and only 10 and 689, respectively, of the common and shared methQTLs associated with a CpG present on the 450k array. We found that most of the common (9/10, Fisher test p-value: 1.6×10^{-4}) and



Enrichment of Tissue-Specific MethQTLs in Proximal Enhancer Elements

To determine characteristic properties of tissue-specific methQTLs, we compared all 452 tissue-specific methQTLs with 1,470 methQTLs shared across multiple tissues. While the distance between the CpG and the SNP that significantly correlates with its DNA methylation state was not different in the two classes of methQTLs, we found both stronger effects with respect to effect size and lower p-values for the shared methQTLs than for the tissue-specific methQTLs (Figure 3.12A). To determine whether the CpGs or the SNPs of the shared and cell-type-specific methQTLs are preferentially located in particular functional regions of the genome, we performed enrichment analyses for various functional annotations such as gene promoters and proximal enhancers. We found that both the SNPs and the CpGs were depleted in regions of active transcription such as transcriptional start sites (TSS) and gene bodies for the shared methQTLs (Figure 3.12B). No significant enrichment towards a functional category was detected for the shared methQTLs. In contrast, the tissue-specific methQTLs were preferentially located in proximal enhancer elements further pointing toward the important regulatory role of enhancers in establishing cellular identity. Further indications for this hypothesis was obtained by the LOLA enrichment of tissue-specific methQTLs toward enhancer elements and transcrip-

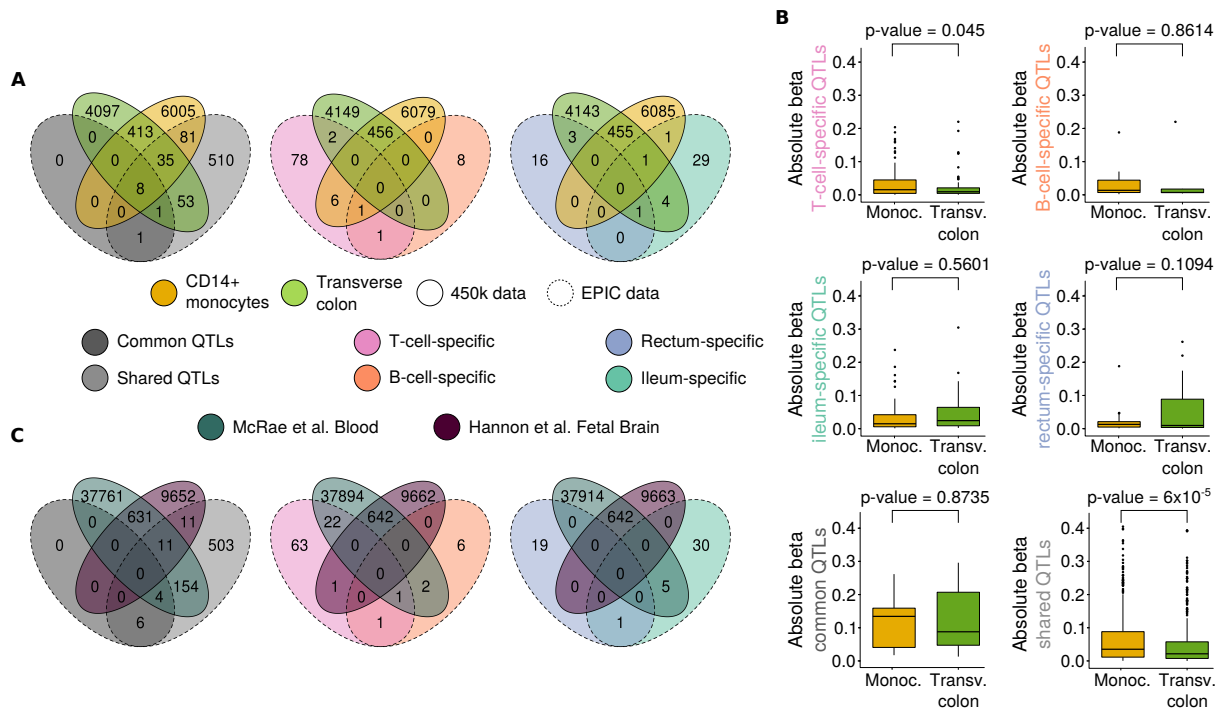


Figure 3.11: Validation of tissue-specific and common methQTLs in the validation cohort and in two independent studies. **A:** Replication of tissue-specific, common, and shared methQTLs in CD14-positive monocytes and transverse colon samples assayed with the Infinium 450k microarray. **B:** Effect size comparison of the T-cell-specific, B-cell-specific, shared, and common methQTLs in CD14-positive monocytes and transverse colon samples. **C:** Replication of tissue-specific, shared, and common methQTLs in published methQTL studies in blood and fetal brain samples.

tion factor binding sites indicating an enhancer element in B-cells and in the B-lymphocyte cell line GM12878 (H4K3me3, Figure 3.12C). Analogously, we associated the tissue-specific and shared methQTL SNPs and CpGs with overlapping gene bodies. For those overlapping genes, we performed Gene Ontology (GO) enrichment analysis and detected an enrichment of the shared methQTLs towards the biological process “cell development” (p-value=0.0069).

We aimed to validate the tissue-specific methQTLs in the validation cohort and in independent studies. While some of the ileum- and rectum-specific methQTLs identified earlier were also present in the transverse colon samples, only two of them were present (at p-value cutoff 9.84×10^{-6}) in the monocytes. Similarly, two of the T-cell-specific methQTLs were also found in transverse colon. However, more (seven for T-cells, one for B-cells) were detected in the CD14-positive monocytes (Figure 3.11A). To validate whether T-cell- and B-cell-specific methQTLs actually capture effects specific to blood cell types, we compared the methQTL effect sizes in the monocytes and in transverse colon. We detected significantly higher effect sizes for the T-cell-specific methQTLs in the monocytes in comparison to transverse colon (Figure 3.11B). Notably, not all methQTLs detected in the discovery cohort could be identified in the validation cohort, since the latter has been assayed using the Infinium 450k technology. Similarly, more of the T- and B-cell-specific methQTLs were present in the methQTL study on blood samples in comparison to fetal brain samples (Figure 3.11C).

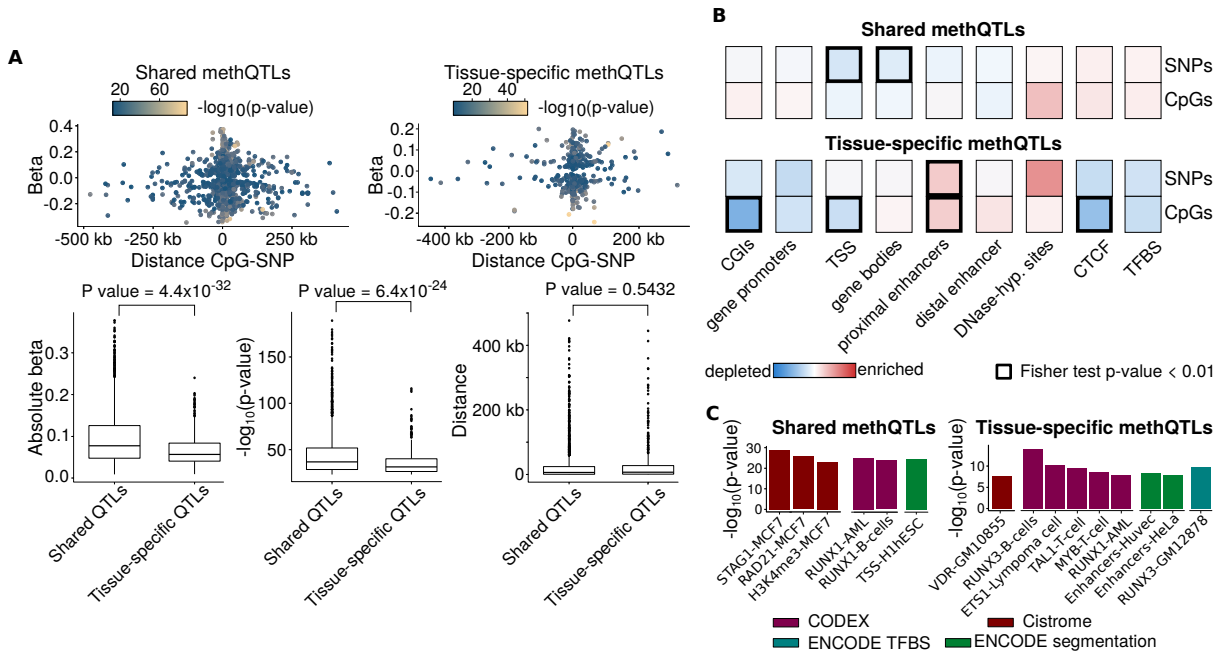


Figure 3.12: Properties of methQTLs shared across different tissues and tissue-specific methQTLs. **A:** Distance between the CpG and the SNP, the effect size of the methQTL, and the negative common logarithm of the methQTL p-value. MethQTLs were classified as either shared or tissue-specific. **B:** Enrichment analysis of shared (upper) or tissue-specific methQTLs (lower) in different functional annotations of the genome. Visualized is the common logarithm of the odds ratio and the associated Fisher exact-test p-value was computed. P-values below 0.01 are indicated by a bold border. **C:** LOLA enrichment analysis of the methQTL SNPs for the shared and tissue-specific methQTLs, respectively.

3.3.4 Discussion

Patient-stratification according to mutational signatures, i.e., genotype-based markers, are already well-accepted in the clinic [213]. At the same time DNA methylation-based biomarkers are also becoming relevant in a clinical setting [85] and may contribute to clinical decision making. The relationship between genotype and DNA methylation variation is only just beginning to be understood. As a first step towards the joint characterization of DNA methylation patterns and genotypes, methylation quantitative trait loci (methQTLs) have been identified in healthy individuals. To facilitate standardized analyses of DNA methylation and genotyping data, we developed the R-package *MAGAR* that supports processing of raw data and integrates with established bioinformatic tools. *MAGAR* is the first package providing a start-to-finish workflow for microarray-based methQTL studies and supports bisulfite sequencing data, without specifically using the additional information present in the sequencing reads. For bisulfite sequencing data, specialized methods are available such as *IMAGE* [214]. Notably, *MAGAR* performs methQTL analysis while accounting for the correlation structure of neighboring CpGs and is a first step toward associating genetic haplotypes with DNA methylation haplotypes. Grouping together CpGs into clusters is an approach that has also been used earlier [215, 216] in contexts different from methQTL analysis. The earlier approaches to group CpGs into correlation blocks however either do not take into account the genomic distance between two CpGs or are restricted to either microarray or bisulfite sequencing data.

It remains elusive whether methQTLs are inherently cell-type-specific or tissue-independent. In this study, we systematically investigated cell-type specificity of methQTLs in sorted blood cell types (CD19+ B-cells, CD4+ T-cells) and bowel biopsies (ileum, rectum). We found fewer tissue-specific methQTLs than methQTLs that were shared across tissues. We validated tissue-specificity in additional CD14+ monocyte and transverse colon samples. Since DNA methylation is a cell-type-specific epigenetic mark, it is likely that methQTLs are also cell-type specific. It remains to be shown whether these cell-type-specific methQTLs preferentially co-occur with other cell-type-specific epigenetic marks such as open chromatin or histone modifications. Previous methQTL studies [57, 208] identified a partially overlapping list of methQTLs, some of which were also detected in this study. Notably, the previous studies used a different strategy for identifying methQTLs (*Merlin* [217] in the blood study and *Matrix-eQTL* [197] in the fetal brain samples). While these strategies do not account for the properties of DNA methylation data, we found a substantial overlap with the methQTLs that we identified.

We found that cell-type-specific methQTLs were preferentially located in enhancer elements, which further emphasizes the importance of enhancers for establishing cellular identity. However, methQTL effects were weaker in cell-type-specific methQTLs compared to those shared across different cell types. It remains to be shown how methQTLs affect gene expression states. In subsequent analyses, the overlap between methQTLs and eQTLs can be explored to further understand the relationship between genome, epigenome, and transcriptome. Since the cell-type-specific methQTLs had weaker effects on the CpG methylation states, cell-type-specific methQTLs could modulate transcript abundance in a more fine-grained manner. We would also like to point out that this observation may be due to technical rather than biological issues.

There are some aspects of methQTLs, which remain to be investigated. It would be relevant to study cell-type-specificity of methQTLs in cell types outside of the hematopoietic system, such as in neurons, epithelial cells, and hepatocytes. To that end, the identified common methQTLs could be further validated to determine whether they are truly tissue- and cell-type-independent. Furthermore, *MAGAR* groups together CpGs into CpG correlation blocks, which reduces the number of SNPs associated with CpGs in the same regulatory unit. However, methQTLs affecting single CpGs may be missed using this method. It is well-established that genetic associations with a disease (GWAS hits) are preferentially located in non-coding regions of the genome [218]. The functional impact of such genetic variants, which can be modulated by QTLs (methQTLs, eQTLs), remains to be investigated. Additionally, DNA methylation data can be used to reliably estimate the cellular proportions of different cell types in the samples using deconvolution analysis (see Chapter 4). Given the cell-type specificity of a subset of methQTLs identified within this study, a combination of DNA methylation-based deconvolution and identification of methQTLs could be implemented similarly to transcriptome-based approaches [195, 219]. By using such a method, it will be possible to investigate methQTL effects in bulk tissues without considering cell-type-specific signals. Preferably, novel analysis methods, such as colocalization analysis and the integration of methQTLs and DNA methylation-based deconvolution, are implemented in an easy-to-use software package such as *MAGAR*. To deal with the issue of cell-type specificity, DNA methylation can be assayed at the single-cell level and associated with genotype information from the same cell. Alternatively, more readily accessible single-cell RNA-seq datasets can be integrated with bulk methQTL studies to understand gene regulation at the single cell level. Finally, long read sequencing allows for simultaneously profiling of the genotype and DNA methylation state of the same molecule

over distances up to 10 kb, which enables associating genetic haplotypes with DNA methylation haplotypes.

In summary, the relationship between genetic and epigenetic variations are currently under-explored. To facilitate the joint analysis of genotype and DNA methylation data, we present *MAGAR* as a novel software tool that accounts for the properties of DNA methylation data. In combination with colocalization analysis, we found tissue-specific and common methQTLs with unique biological properties and genomic location. The identified tissue-specific and common methQTLs were also present in independent samples.

DNA Methylation Heterogeneity Between Samples Sharing a Phenotype

DNA methylation varies between phenotypes (e.g., diseased versus healthy individuals) as discussed in the previous chapter, but it also exhibits substantial variation within a phenotypic group. To comprehensively address heterogeneity within a group, which is mainly driven by cell-type heterogeneity, computational deconvolution tools such as MeDeCom (Lutsik et al. [220]) have been developed. These deconvolution methods require data preprocessing and biological interpretation of the results. In a collaborative project with Pavlo Lutsik and Reka Toth from DKFZ Heidelberg, and Petr V. Nazarov from the LIH in Luxembourg among others, I developed a three-stage protocol for performing deconvolution of complex DNA methylomes. The collaboration was established at the Health Data Challenge in Aussois, France, in 2018, and the major findings of the challenge have resulted in the publication Decamps et al. [221]. I mainly contributed to the development of the first (preprocessing) and last (biological interpretation) stages of the protocol. Additionally, I applied the presented protocol to cancer data from TCGA. The first part of this chapter is a modified version of the manuscript Scherer et al. [222] published in Nature Protocols (2020). Here, I merely give an overview and do not discuss the full step-by-step protocol as presented in the original publication.

In the second part of the chapter, I report on an application of the deconvolution pipeline to DNA methylation data from melanoma patients. In a project led by Katharina Filipinski, Kim Zeiner, Pia Zeiner, and Patrick Harter, I conducted deconvolution analysis of melanoma metastases of patients treated with immune checkpoint inhibition (ICI) therapy. The identified components showed an association with immune cell infiltration and with patient survival after ICI treatment. A manuscript describing the analysis has been submitted.

4.1 Reference-Free Deconvolution, Visualization, and Interpretation of Complex DNA Methylation Data Using *DecompPipeline*, *MeDeCom*, and *FactorViz*

4.1.1 Deconvolution of Complex DNA Methylation Data

In Chapter 3 of this thesis, we reported on methods and analysis tools for comparing DNA methylation states between phenotypes in EWAS. Such studies have been performed to associate CpG methylation states with various diseases and traits, including cancer [76, 84, 78], inflammatory diseases [70], and aging [176]. As a special case of a trait that associates with the DNA methylation state of individual CpGs, we investigated genotypes and their effects on

DNA methylation in methQTL. Notably, EWAS and methQTL studies are mainly performed on bulk tissue samples, such as whole blood or complex biopsies, while DNA methylomes obtained from bulk samples are intrinsically heterogeneous due to different cell types contributing different DNA methylation patterns. DNA methylomes can additionally be affected by age, sex, and technical confounding. To systematically investigate such complex, bulk samples, major sources of variation can be detected from the data and associated with biological properties such as cell-type identity or organism age. To that end, computational methods for the dissection of the between-sample heterogeneity of large-scale DNA methylation datasets into biologically distinct components of variability have been developed, which are of paramount importance for the analysis of bulk samples [144].

Computational deconvolution methods have been developed that separate bulk methylomes into their basic constituents, which we refer to as latent methylation components (LMCs) [223]. These methods can be divided into four classes: *reference-based*, *confounding factor analysis*, *semi-reference-free*, and (fully) *reference-free* methods (Table 4.1). Reference-based methods require DNA methylomes of purified cell types and infer the proportions of these cell types across the samples. Large international consortia, including IHEC, DEEP, and BLUEPRINT, are generating such genome-wide DNA methylation profiles of primary tissue samples and isolated cell populations and facilitate reference-based deconvolution. Multiple reference-based methods have been proposed [224, 225, 226, 227, 228] and are reviewed elsewhere [229]. Most reference-based methods use modifications of linear least squares regression (e.g., robust partial correlations, constrained projection) to infer the proportions. Notably, reference-based methods require selecting cell-type-specific CpGs using pairwise differential analysis, and dedicated methods (cell-type-marker selection methods) have been developed to optimize the selection of cell-type-specific CpGs [229]. The second class of deconvolution methods aims at removing the effect of cell-type heterogeneity from EWAS without explicitly computing the cell-type proportions [230, 231]. For instance, such methods employ PCA on the DNA methylation data restricted to cell-type-specific CpGs. Then, they use the first principal component as a confounding factor (covariate) in the differential analysis and compute DMRs/DMCs that are independent of the cell type. When reference methylomes and other prior information is partially or completely absent, semi-reference-free [232] or fully reference-free [233, 234, 220, 235, 236] deconvolution methods can be applied (Table 4.1). While semi-reference-free methods require some information about the potential cell types present in a sample such as cell-type-specific markers, completely reference-free methods do not require prior information. Reference-free methods use matrix factorization methods, such as non-negative matrix factorization (NMF) to dissect the input DNA methylation data matrix into two matrices: a matrix of CpG methylation states across LMCs and a matrix of LMC proportions across the samples. Various modifications of the approach have been implemented and resulted in software packages such as *RefFreeCellMix*, *EDec*, or *MeDeCom*. In this work, we particularly focus on *MeDeCom* as the deconvolution tool, which employs a regularized version of NMF.

Reference-free deconvolution is particularly useful for dissecting DNA methylomes of biological systems with limited prior knowledge about their cellular composition, or in case reference profiles of purified cell types in a bulk sample are missing. Examples for such biological systems include difficult-to-access or insufficiently characterized organs and tissues, including human brain, as well as solid tumors. Reference-free deconvolution has been employed to understand cellular heterogeneity in placenta [239], multiple sclerosis [240], breast cancer [234],

Table 4.1: Overview of published deconvolution tools using DNA methylation data. The methods are stratified into four classes of deconvolution methods and ordered chronologically according to their date of publication. (Ref. = Reference number)

Tool	Class	Short description	Ref.
<i>Houseman</i>	reference-based	The method employs constrained projection to infer proportions of reference profiles and was particularly developed for deconvolution of whole blood samples.	224
<i>EpiDISH</i>	reference-based	<i>EpiDISH</i> is a reference-based method using robust partial correlations to compute proportions of reference profiles. The authors propose a method based on DNaseI-hypersensitive sites to determine appropriate reference profiles.	225
<i>hEpiDISH</i>	reference-based	<i>hEpiDISH</i> is an extension of <i>EpiDISH</i> that hierarchically performs deconvolution, and together with a new reference database, improves deconvolution results in comparison to <i>EpiDISH</i> .	226
<i>Methyl-CIBERSORT</i>	reference-based	An extension of <i>CIBERSORT</i> [237] created for RNA-seq data that employs support vector regression to estimate the proportions of given DNA methylation reference profiles across the samples.	227
<i>methylCC</i>	reference-based	<i>methylCC</i> uses latent components and a region-based rather than an individual CpG-based model to compute the proportions of given reference profiles independent of the technology used.	227
<i>FaST-LMM-EWASher</i>	confounding factor in EWAS	The <i>EWASher</i> approach is based on linear mixed models to account for differences in cellular compositions in EWAS.	230
<i>ReFACTor</i>	confounding factor in EWAS	<i>ReFACTor</i> uses PCA on sites that are differentially methylated between cell types. The first few principal components are then used to adjust for cell-type composition differences in EWAS.	231
<i>BayesCCE</i>	semi-reference-free	<i>BayesCCE</i> is a semi-supervised method to estimate proportions of different cell types that requires some prior knowledge on the cell-type composition of the studied tissue.	232
<i>RefFree-CellMix</i>	reference-free	<i>RefFreeCellMix</i> from <i>RefFreeEWAS</i> [238] uses NMF of the input DNA methylation matrix to compute a matrix of proportions and estimated reference profiles (LMCs).	233
<i>EDec</i>	reference-free	<i>EDec</i> is a two-step approach that combines reference-based and reference-free estimations using constrained matrix factorization.	234
<i>MeDeCom</i>	reference-free	<i>MeDeCom</i> uses regularized NMF on the input DNA methylation data to create a matrix of proportions and a matrix of LMCs.	220
<i>TCA</i>	reference-free	<i>TCA</i> uses tensor composition analysis to obtain sample-specific cell-type-profile estimates. In contrast to standard NMF, the method returns multiple, sample-specific LMC matrices using the same proportions matrix.	235
<i>CONFINED</i>	reference-free	<i>CONFINED</i> requires two DNA methylation matrices as input and uses canonical correlation analysis to obtain purely biological sources of variation.	236

and cholangiocarcinoma samples [241]. Since tumors are highly heterogeneous, reference-free deconvolution methods are especially useful for dissecting cancer samples. For example, deconvolution analysis can be used to study the effect of tumor-infiltrating immune cells on the tumor microenvironment [242]. Furthermore, the proportions of identified LMCs across the samples can correlate with tumor size, location, metastasis state, and mutational burden, and

thus be informative for patient survival. In cancer studies, methylome deconvolution can detect similarities among different types of cancers and contribute to the discovery of cancer-type-specific DNA methylation patterns and those shared across different cancer types. Since reference-based approaches assume that the constituting cell types of a sample are known, rare cell types are likely to be missed by reference-based, but not by reference-free methods. Similarly, cancers that induce changes in the DNA methylation pattern of the tumor stroma and thus alter the adjacent cell types cannot be investigated by reference-based methods.

Deconvolution methods require input data of high technical quality to obtain the desired major components of variation in the DNA methylome [221]. Furthermore, interpretation of the detected components is challenging in absence of information about the investigated biological system. Thus, we developed a comprehensive pipeline that facilitates reference-free deconvolution, starting from raw DNA methylation data down to result interpretation. Although we focus on *MeDeCom* [220] as a representative method, the protocol is not limited to a single deconvolution strategy and can be used in combination with other available tools including *RefFreeCellMix* [233] and *EDec* [234]. To show the applicability of the protocol, we applied it to lungadenocarcinoma data from The Cancer Genome Atlas (TCGA) and detected components associated with tumor-infiltrating immune cells and prognostic outcome.

4.1.2 A Pipeline for Reference-Free Deconvolution of DNA Methylation Data

Overview

Reference-free deconvolution of DNA methylation data represents a challenging computational problem, for which different approaches have been developed (Table 4.1). In a benchmark of three such deconvolution tools (*MeDeCom*, *RefFreeCellMix*, and *EDec*), we found that results obtained both on fully synthetic and *in-silico* mixed experimental datasets were largely similar [221]. Additionally, we found that the quality and information content of the input DNA methylation matrix was more important for the success of deconvolution than the tool itself, which is an observation made across many problems in Computational Biology. We concluded that deconvolution algorithms require high-quality data, which is obtained through extensive data preprocessing and feature selection. This becomes especially important if the differences between underlying components are small.

Deconvolution tools return a matrix of LMCs and a matrix of LMC proportions across the samples. It is especially critical to assign biological properties to those matrices, which is particularly challenging for beginners with limited experience with the analysis of biological data. In order to facilitate biological interpretation of reference-free deconvolution results, we developed an R/shiny application. The summary of these steps resulted in a three-stage protocol for analyzing complex DNA methylation datasets.

The three stages of the protocol, schematically outlined in Figure 4.1, are: (i) data preprocessing, (ii) deconvolution, and (iii) interpretation.

1. **Data preprocessing** is crucial for the overall success of deconvolution, since deconvolution tools require input data of high technical quality. To facilitate the generation of such a high-quality dataset, the first stage of our protocol (*DecompPipeline*) uses *RnBeads* (see also Chapter 3) for handling DNA methylation data. Potentially unreliable CpGs are removed from the analysis. Since age, sex, or the genetic background can substantially

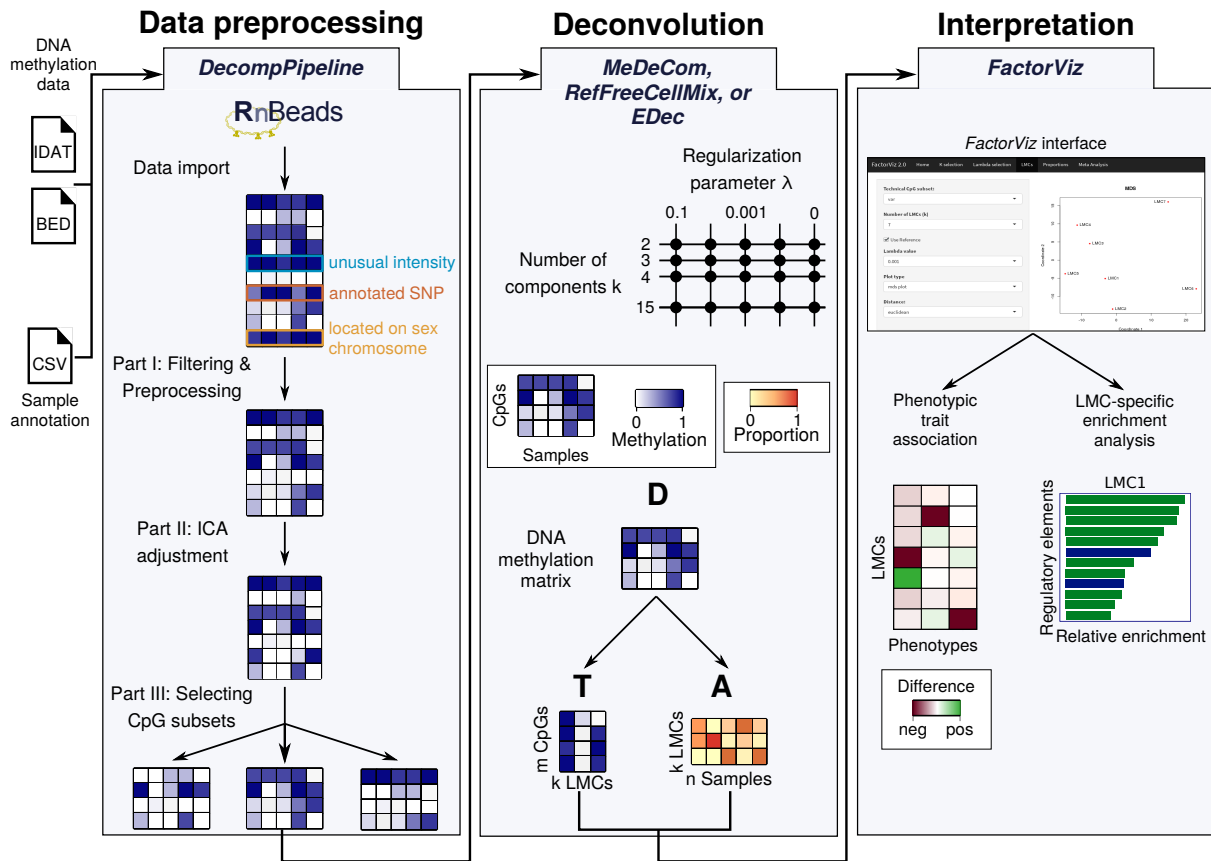


Figure 4.1: Overview of the proposed deconvolution protocol. DNA methylation data can be used from any technology yielding single-CpG methylation calls. Methylation data is first processed using *DecompPipeline*, which includes data import, preprocessing, accounting for confounders and feature selection. *MeDeCom*, *RefFreeCellMix*, or *EDec* can be used to perform deconvolution of the input methylation matrix (dimension m CpGs \times n samples) into the latent methylation components (LMCs) and the proportions matrix (dimension k LMCs \times n samples). A grid of values for the regularization parameter λ and the number of components k has to be specified. The resulting matrices are then validated and interpreted using the R/Shiny visualization tool *FactorViz*.

affect the methylome [45, 44], these factors are typically considered as confounding factors. Within the protocol, we propose independent component analysis (ICA) [243] to account for confounding factors. In the presented example analysis, we argue that this adjustment is crucial for obtaining biologically relevant results. Since only few CpGs will contribute to the discovery of LMCs, we select a subset of CpGs that are associated with features such as cell-type identity or any other phenotypic trait of interest as the final step of the preprocessing.

2. The processed DNA methylation data matrix is used as input to one of the reference-free **deconvolution** tools *MeDeCom*, *RefFreeCellMix*, or *EDec*. These methods decompose the input DNA methylation matrix into the LMC matrix T and a matrix of proportions of LMCs across the samples (A). Detected LMCs correspond to major sources of variation in the methylome including, but not limited to, DNA methylation profiles of underlying

cell types. The proportions matrix A quantifies the relative contribution of each LMC to each (bulk) sample. In the use case that we discuss here, we use *MeDeCom* [220], a method based on regularized NMF for deconvolution, but the proposed protocol seamlessly integrates with other deconvolution tools.

3. Reference-free deconvolution can be applied to any biological system, which introduces challenges for the **interpretation** of deconvolution results. Most notably, the LMC matrix T has to be scanned for unique biological properties. The detected LMCs reflect multiple drivers of biological and technical variability, including cell-type composition. Furthermore, technical and biological validation of the proportions and LMCs is not trivial, since the underlying ground truth (e.g., the cellular composition) is typically unknown. For the interpretation of the detected components, we propose tests of association to available sample metadata and enrichment analysis of LMC-specific CpGs. To make the various validation and interpretation functions available also to users with limited bioinformatic knowledge, we developed the specialized R/Shiny-based graphical user interface *FactorViz* for the interpretation of deconvolution results.

The protocol that we present below is available as a step-by-step procedure on the Supplementary Website¹. Additionally, the software packages *DecompPipeline*², *MeDeCom*³, and *FactorViz*⁴ are freely available from GitHub.

Data Preprocessing

We compiled the data preprocessing steps required for successfully performing reference-free deconvolution of DNA methylation data as a new R-package (*DecompPipeline*). *DecompPipeline* integrates quality filtering, adjustment for confounding factors, and feature selection into a comprehensive workflow. For loading, formatting, and storing DNA methylation data we recommend *RnBeads* (see also Chapter 3).

Data Import Genome-wide DNA methylation profiles can be obtained using different technologies such as WGBS, RRBS, or the Illumina microarrays [90]. Within this project, we used 450k microarray data, which is still the technology used for most publicly available datasets assaying DNA methylation. Notably, deconvolution results tend to improve with higher number of samples. However, the protocol can be applied to Illumina EPIC data after adjusting some of the steps such as selecting an appropriate normalization technique (see Section 4.2 for an application to EPIC data). Additionally and similar to *RnBeads*, our pipeline is applicable to any other data type that provides DNA methylation calls at single CpG resolution, including RRBS and WGBS. Some of the steps have to be modified for bisulfite sequencing data, and *DecompPipeline* provides specific functions for processing WGBS/RRBS data. These steps include merging the CpGs assayed in different samples, and read coverage filtering of CpGs and samples. In addition to raw DNA methylation data, phenotypic information, such as donor age and sex, is required and converted into the internal *RnBeads* data structures. We use *RnBeads*' QC

¹<http://epigenomics.dkfz.de/DecompProtocol/>

²<https://github.com/CompEpigen/DecompPipeline>

³<https://github.com/lutsik/MeDeCom>

⁴<https://github.com/CompEpigen/FactorViz>

Table 4.2: Configurable quality filtering options available in *DecompPipeline*.

450k/EPIC array and bisulfite sequencing	
Missing value filtering	Removes sites comprising missing measurements in any of the samples
SNP filtering	Removes sites or probes overlapping with SNPs (minor allele frequency > 1%, dbSNP147 [113])
Sex chromosome filtering	Filters sites located on the sex chromosomes
450k/EPIC array only	
Bead filtering	Filters sites covered by less than <code>min.n.beads</code> (default=3) beads in any of the samples
Intensity filtering	Filters sites according to their overall intensity values with respect to the intensity quantiles of all sites specified using <code>min.int.quant</code> (default=0.001) and <code>max.int.quant</code> (default=0.999)
Cross-reactive filtering	Removing sites reported to be cross-reactive [114, 91]
Bisulfite sequencing only	
Absolute coverage filtering	Filters sites with read coverage less than <code>min.coverage</code> (default=5)
Quantile coverage filtering	Removes sites according to the read coverage quantiles (defined by <code>min.covg.quant</code> and <code>max.covg.quant</code>)

module to check the raw data for quality. Preprocessed data can also directly be used as an input for the pipeline as a DNA methylation data matrix, which enables integration with further DNA methylation processing tools, such as *minfi* [111], *wateRmelon* [116], or *ChAMP* [160].

Quality Filtering and Covariate Inference By applying *MeDeCom* to multiple datasets, we found that deconvolution analysis is especially sensitive toward technical batch effects. Thus, we use very stringent quality criteria and a step-wise approach for focusing on a smaller subset of CpGs in *DecompPipeline* (Table 4.2). First, we filter CpGs according to a bead or read coverage threshold across the samples, i.e., a CpG must fulfill the threshold in all of the samples. Second, CpGs that show unusually high or low signal intensity (microarrays) or read coverage (bisulfite sequencing) are removed. Since missing values are not accepted in further steps of the protocol, they can either be completely discarded from the dataset or imputed (see Section 3.1.2). To avoid confounding by genetic background of the samples, sex, and cross hybridization, we further remove sites overlapping annotated or estimated SNPs [113], sites on the sex chromosomes, and cross-reactive sites ([114, 91], cf. Section 2.5.1). Infinium data can be normalized prior to downstream analysis [93, 117, 94] to account for the bias resulting from the design of the microarrays. Lastly, additional sample properties can be inferred including the overall immune cell content using the LUMP algorithm [149] or the epigenetic age [44]. Reasonable default values were selected based on our experience with the analysis of DNA methylation data (Table 4.2).

Covariate Adjustment with ICA DNA methylomes are affected by various sources of variability including biological and technical influences that might mask the signals of interest. It is a critical choice to decide which of the detected components are of relevance for the investigated system and which components are associated with unwanted sources of variation (confounding factors). For instance, components associated with age may be relevant for studying age-related phenotypes, while age is a notable confounding factor for tumor heterogeneity.

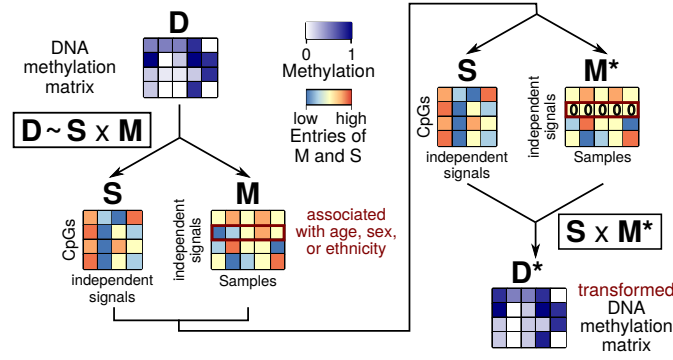


Figure 4.2: Overview of covariate adjustment using ICA. The input DNA methylation matrix is decomposed into two matrices (S) and (M). Components (rows of M) associated with confounding factors (here sex, age, or ethnicity) are removed from the contribution matrix (M) and an adjusted DNA methylation matrix (D^*) is constructed.

To allow for flexibility, we leave it to the user to decide which factors are to be considered as confounding, but recommend to use age, sex, and ethnicity as default values. Additionally, in case different batches of samples were used, this information should also be included. Within our protocol, we propose to use ICA [243] as a data-driven dimensionality reduction method that performs a matrix decomposition for adjusting the data according to a given set of confounders. Analogously to NMF, ICA divides the experimentally observed data matrix $D_{p \times n}$ into k independent signals $S_{p \times k}$ mixed with the coefficients of $M_{k \times n}$:

$$D_{p \times n} \sim S_{p \times k} \times M_{k \times n} \quad (4.1)$$

where p and n are the number of CpGs and samples, respectively. ICA does not impose restrictions on the entries of the matrices in contrast to *MeDeCom*, which only allows entries in the $[0, 1]$ interval. Notably, the entries of the LMC matrix can be considered as DNA methylation values for *MeDeCom*, while this is not a requirement for ICA. To determine associations between Independent Components (ICs) and a set of confounders, the weight matrix M can be associated with a set of potential confounding factors, such as age or sex. Similarly, the statistically independent signals can be attributed to individual CpGs. We offer two choices to account for the effect of a confounding factor: Either the CpGs associated with the confounder can be removed from the analysis, or the weights (entries of M) can be set to zero [244]. The latter method preserves all CpGs for the analysis, but modifies the DNA methylation matrix. If the influences of the investigated confounding factors are small, we recommend setting the corresponding components to zero and reconstructing an adjusted data matrix (Figure 4.2). We apply the consensus ICA approach to obtain the matrix decomposition [245]. The integration of ICA into *DecompPipeline* has been realized by Tony Kaoma and Petr Nazarov from the LIH in Luxembourg.

Selection of Informative CpGs Feature selection is another important step of the pipeline, since, for instance, lowly variable CpGs do not contribute to the identification of LMCs, but add to the computational runtime. Additionally, considering CpGs not associated with the outcome of interest can mask more subtle signals. From our experience, integrating prior knowledge about the underlying cell types, for instance known cell-type-specific CpGs, is the best option.

Table 4.3: CpG selection options available in *DecompPipeline*. recom.=recommended

CpG selection method	CpG subset selected	Details	450k	EPIC	sequencing
VAR	Most variable across the samples	n.markers to	✓	✓	✓
RANDOM	Random subset	determine the	✓	✓	✓
HYBRID	Half most variable, half randomly	number of sites	✓	✓	✓
PCA	Highest loadings on the first	Default=10	✓	✓	✓
PCADAPT	n.prin.comp principal components PCA implemented in the <i>bigstatsr</i> R-package	Privé et al. [203]	✓	✓	✓
ALL	All that fulfill the quality criteria		✓	✓	Not recom.
RANGE	Largest dynamic range across the samples		✓	✓	✓
CUSTOM	User-specified list		✓	✓	✓ (recom.)
ROWFSTAT	Linked to given reference profiles using the F-statistics	Requires reference profiles	✓	✓	✓
PHENO	Differentially methylated according to specified phenotypic groups using the <i>limma</i> method	Ritchie et al. [246]	✓	✓	✓
HOUSEMAN2012	50,000 sites determined to be cell-type-specific using the Houseman et al. method and the Reinus et al. reference dataset, applicable only to blood datasets	Houseman et al. [224], Reinus et al. [247]	✓	✗	✗
HOUSEMAN2014	According to the <i>RefFreeEWAS</i> method	Houseman et al. [238]	✓	✓	✓
JAFFE2014	600 sites listed as cell-type-specific in Jaffe et al., applicable only to blood datasets	Jaffe et al. [248]	✓	✗	✗
EDEC_STAGE0	According to Stage 0 of the <i>EDec</i> approach; requires reference profiles	Onuchic et al. [234]	✓	✓	✗

Such knowledge is typically available for well-characterized systems, such as whole blood [224, 234]. In the absence of prior knowledge about the biological system of interest, the protocol provides multiple feature selection methods, including selecting the most highly variable sites, the ones with the highest loadings on the first few principal components, or a random selection of sites. In total, *DecompPipeline* provides 14 such options (Table 4.3), and multiple of these options can be included in a single execution of the pipeline.

Performing Deconvolution using *MeDeCom*

Reference-free deconvolution tools, such as *RefFreeCellMix* [233], *EDec* [234], or *MeDeCom* [220], dissect the DNA methylation matrix ($D_{p \times n}$) of sites selected in the previous step into the LMC matrix ($T_{p \times k}$) and their proportions across the samples ($A_{k \times n}$).

$$D_{p \times n} \sim T_{p \times k} \times A_{k \times n} \quad (4.2)$$

For the tools stated above, non-negative matrix factorization (NMF) is at the core of deconvolution, and the different tools solve modified version of the NMF problem. *MeDeCom* is used for the presented analysis, but the pipeline similarly supports *RefFreeCellMix* and *EDec*. Notably, the deconvolution itself is independent of the input data type used, since a DNA methylation data matrix can either be generated through RRBS/WGBS or using the Illumina microarrays. *MeDeCom*'s objective is the minimization of the squared Frobenius norm of the difference between the true (observed) methylation matrix D and the matrix product of T and A (Equation 4.3). The Frobenius norm of a matrix is the square root of the sum of its squared entries. Additional constraints ensure that the estimated matrices are restricted to entries in the $[0, 1]$ interval (T and A), and that the column sums have to be equal to one for A . The original motivation behind these constraints is that the entries of T can be interpreted as DNA methylation values for CpGs across the LMCs. Additionally, the columns of A should sum up to one in order to interpret the entries as LMC proportions across the samples. A special modification employed by *MeDeCom* is the penalization of the entries of T not equal to zero or one using quadratic regularization (maximal at entries equal to 0.5). To control for the strength of regularization, the hyperparameter λ was introduced. In summary, the computational problem solved by *MeDeCom* can be formulated as follows:

$$\begin{aligned} \text{Objective:} \quad & \min_{T,A} (\|D - TA\|_F)^2 + \lambda \sum_{i=1}^m \sum_{s=1}^k T_{is}(1 - T_{is}) \\ \text{subject to} \quad & 0 \leq T_{is} \leq 1, \forall i, s \\ & A_{sj} \geq 0, \forall s, j \\ & \sum_{s=1}^k A_{sj} = 1, \forall j \end{aligned} \quad (4.3)$$

To find an optimal solution for the problem, *MeDeCom* employs an alternating optimization scheme. This means that at each step of the algorithm, the quadratic optimization problem is solved for either A or T while the other matrix is kept fixed. Hyperparameter selection for the regularization parameter (λ) and the number of latent components (k) is realized through a cross-validation scheme that leaves out columns of D and computes the reconstruction error (referred to as the cross-validation error). The cross-validation error has been implemented, since the objective value will always decrease with higher numbers of k . Typically, a grid of different values of k and λ is specified, and the user needs to determine the most suitable number of components and regularization, respectively, using the diagnostic plots returned by *MeDeCom*. In order to reduce runtime substantially, we recommend activating the parallel processing options on standalone workstations, or to use an HPC cluster. To facilitate downstream analysis, the deconvolution results are stored as internal data structures, which serve as input to *FactorViz* and can be investigated using custom R scripts⁵.

Interpretation of Deconvolution Results

For reference-based deconvolution methods, the output is a matrix of proportions of given cell types across the samples. These proportions can be associated with a phenotype of interest,

⁵See e.g., <http://epigenomics.dkfz.de/DecompProtocol/resources.html>

checked for biological plausibility, or be used as covariates in differential analysis (cf. Chapter 3). In contrast, *MeDeCom* returns two matrices: a LMC matrix and a matched proportions matrix. Both matrices need to be biologically validated and interpreted, and we created the semi-automated visualization tool *FactorViz* to facilitate the visualization and interpretation of deconvolution results. Notably, *FactorViz* also accepts *RefFreeCellMix*'s and *EDec*'s output. *FactorViz* has been implemented as an R/Shiny-based user interface, and provides guidelines and functions for comprehensive biological inference.

As a first step of the analysis, the user selects one of the *MeDeCom* solutions by determining reasonable values for the parameters k and λ based on the cross-validation error. We recommend selecting the values of k and λ that lead to a small cross-validation error, while other statistics such as the objective value of the optimization problem are also small. Notably, the cross-validation error tends to decrease when more components are considered, and we recommend selecting the value of k at which the cross-validation error starts to level off. After fixing a value for k , a reasonable value for λ can be determined by plotting the cross-validation error and the objective value for the different values of λ . An optimal value of λ is located at the minimum of both error metrics, but often one has to trade off between the cross-validation error and the objective value (see also Figure 4.5). To determine potential influences of covariates upon the estimated proportions and corresponding LMCs, the proportion matrix returned by *MeDeCom* is linked to technical or phenotypic traits using association tests such as two-sided correlation and t-tests. Furthermore, in case matched gene expression profiles are available, proportions can be associated with expression of known cell-type marker genes. In application to cancer datasets, the LMCs can be related to cancer-specific properties such as survival time [249] or cancer subtype. To determine clinically distinct sample subgroups, the proportions matrix can be used to obtain sample groups using hierarchical or k-means clustering.

For functional annotation of LMCs, we determine the CpGs that are specifically hypomethylated in an LMC in comparison to the median of the remaining LMCs and call the obtained sites LMC-specific. These LMC-specific CpGs are used for GO [250] and LOLA [152] enrichment analysis in order to associate respective LMCs with functional categories, pathways, and genomic features. Finally, in case reference profiles are available for a subset of the cell types present in the samples, the LMC matrix can be compared to those profiles. Notably, interpretation of deconvolution results is independent of the technology used, and RRBS/WGBS data can be similarly analyzed using *FactorViz*. *FactorViz* has originally been developed by Pavlo Lutsik and has been improved by Shashwat Sahay.

4.1.3 Reference-Free Deconvolution of Lung Adenocarcinoma Data

To show a use case of the presented pipeline, we collected publicly available data from TCGA⁶ investigating lung adenocarcinoma (dataset TCGA-LUAD [251]⁷) in 461 samples assayed with the Illumina 450k microarray. We used this dataset, since lung cancer is characterized by high cellular and molecular heterogeneity [252]. The dataset comprises cancer samples, as well as samples from cancer-adjacent, healthy tissue. We employed the presented protocol and used default parameters for most of the analysis steps, unless stated otherwise. Notably, we accounted for age, sex, and ethnicity as potential confounding factors and selected the 5,000 most

⁶<https://www.cancer.gov/tcga>

⁷<https://portal.gdc.cancer.gov/projects/TCGA-LUAD>

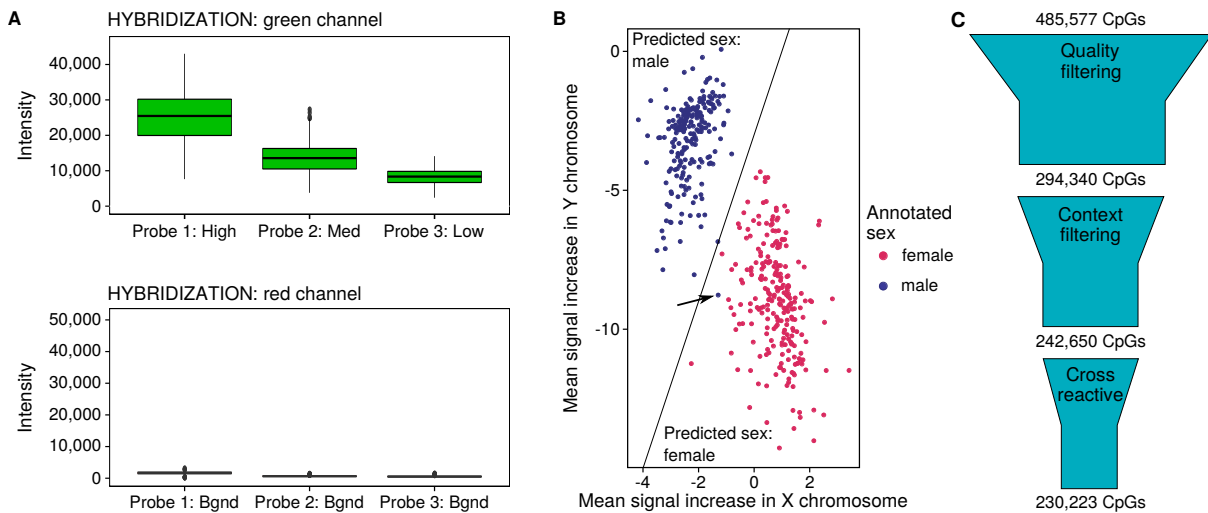


Figure 4.3: Quality control of lung adenocarcinoma data from TCGA. **A:** Boxplot for signal intensities of hybridization control probes for the green and the red channel, respectively. **B:** Sex prediction based on the intensities of the probes on the sex chromosomes. The values on the axes indicate the relative signal intensities on the sex chromosomes in comparison to the autosomes. The decision boundary of *RnBeads*' logistic regression classifier to differentiate between female and male samples is indicated, and an incorrectly classified sample is indicated by an arrow. **C:** Outline of the CpG filtering procedure. All CpGs on the 450k array are filtered according to quality scores (coverage, overall intensity) and genomic sequence context (SNPs, sex chromosomes), and cross-reactive sites are discarded.

variably methylated CpGs as input for the deconvolution.

Quality Control and Feature Selection

Since reference-free deconvolution analysis requires input DNA methylation data of high technical quality, we verified data quality using *RnBeads*' QC module (see also Chapter 3). All quality control probes on the Infinium microarray showed the expected distribution of signal intensities, i.e., the high, medium, and low intensity control probes showed substantially higher signal intensities than the background control probes. In addition to validating hybridization (Figure 4.3), further control probes are available, for instance for bisulfite conversion, specificity, or extension, which also showed the expected distribution across the samples (see Supplementary Figure A.5 for an example of bad data quality). Additionally, annotated phenotypic information matched the inferred sample properties, such as predicted sex for all but one of the subjects (Figure 4.3). This is likely an incorrect prediction of the *RnBeads*' sex classifier, since the sample did not exhibit any other unusual behavior. Thus, no sample was removed from downstream analysis.

In order to select a set of high-confidence CpGs as input to *MeDeCom*, we employed various, stringent inclusion criteria. First, a large fraction of CpGs (39.5%) was removed, because they were covered by fewer than three beads (cf. Section 2.4) in any of the samples, or showed unusually high or low signal intensities. These sites are potentially problematic, since low coverage might induce spurious associations due to technical rather than biological reasons. Furthermore, only a small subset of CpGs is required for component detection by *MeDeCom*. Intensity

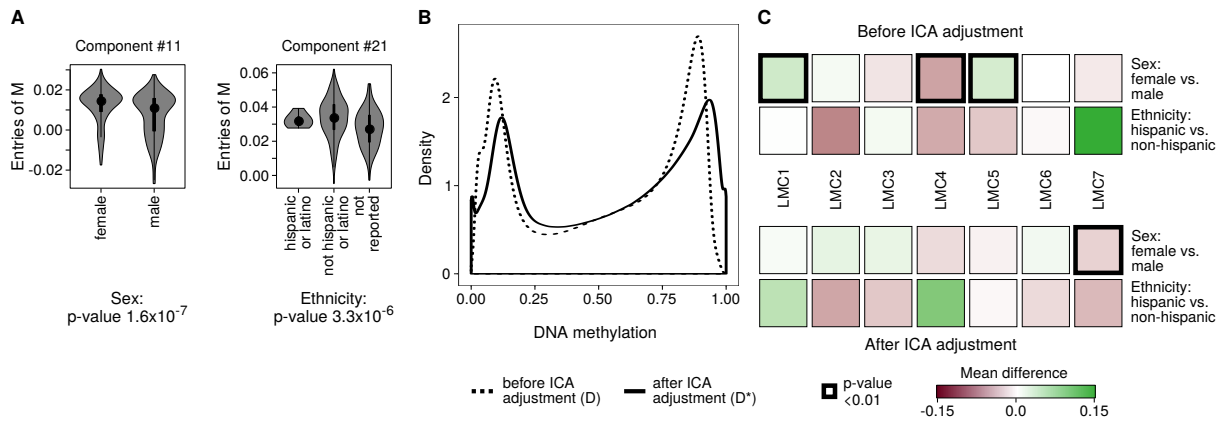


Figure 4.4: Evaluation of ICA on lung adenocarcinoma data. **A:** Associations between the confounding factors sex and ethnicity with the entries of the proportion matrix M returned by ICA. **B:** Beta-value distributions of the DNA methylation data matrix after ICA-based transformation (D^*) and of the untransformed matrix (D). **C:** Associations between LMC proportions generated in two independent *MeDeCom* runs (using either D^* or D as input) and qualitative phenotypic traits. The colors represent the absolute difference of the mean LMC proportions in the different groups defined by sex/ethnicity and significant p-values according to a two-sided t-test are indicated by a bold outline.

outliers, on the other hand, can be main drivers of the LMCs, which is an undesirable outcome. Additional CpG filtering steps included sequence context filtering (SNPs, sites on the sex chromosomes, 10.5%) and removal of potentially cross-reactive probes (2.5%) [114, 91]. Finally, 230,223 sites were retained (47.4% of 485,577) after stringent quality filtering and used for downstream analysis. Notably, these extremely stringent quality criteria might need to be adapted to each dataset individually, but the input to *MeDeCom* should be selected from a small set of highly reliable CpGs.

Confounding Factor Analysis

We evaluated the proposed covariate adjustment using ICA by applying the presented workflow to the TCGA dataset twice: once without adjusting for age, sex, and ethnicity, and once with the adjustment using ICA. ICA revealed 22 independent components, of which two were significantly associated with sex and ethnicity, respectively (Figure 4.4A). Notably, we set the components of ICs 11 and 21 in M to zero in this application, since we expect that the confounding factors only mildly influence the DNA methylation data. A comparison between the overall data distribution of the adjusted and non-adjusted DNA methylation matrix revealed that the transformed DNA methylation data matrix still showed the expected bimodal distribution after ICA adjustment (Figure 4.4B). The pipeline automatically returns the plots shown in Figure 4.4, and the effect of ICA-based adjustment should be carefully investigated. In case the effect of the ICA-based reconstruction of the data matrix (see Figure 4.2) is stronger than described here, users can decide to remove the CpGs affected by confounding factors from the original data matrix rather than adjusting the data matrix.

To show that the proposed ICA-based adjustment successfully removes the confounding factor from the data, we executed *MeDeCom* independently on the unadjusted and the adjusted DNA methylation matrix. As a result, three of the detected components were significantly as-

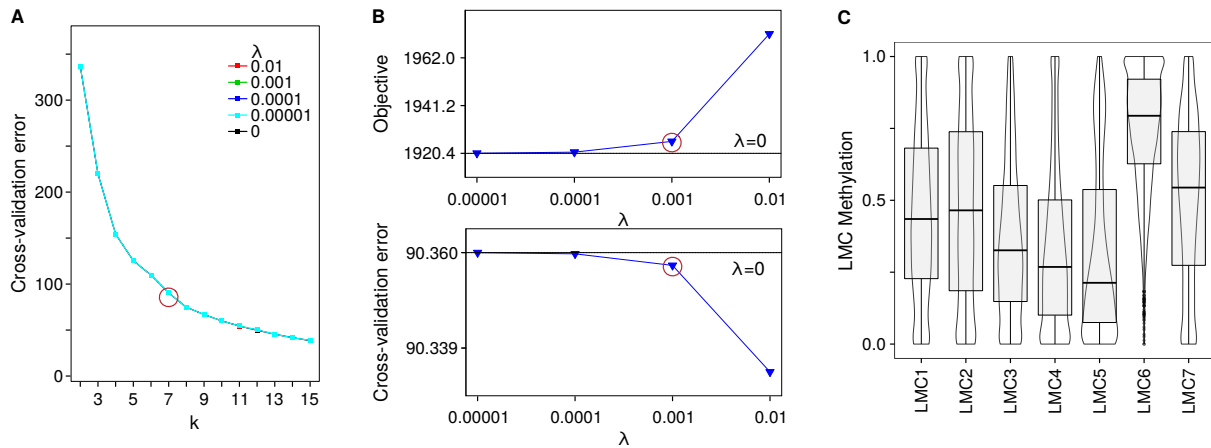


Figure 4.5: Selecting the number of LMCs and the regularization parameter for *MeDeCom*. **A:** Cross-validation error plotted against the number of LMCs k for different values of the regularization parameter λ . Differences across the range of values for k mask the differences between the five λ values used. **B:** Objective value and cross-validation error for different values of λ after fixing the number of components to seven. **C:** Combined violin- and box-plots of the LMC methylation matrix for the selected parameters.

sociated with sex in the unadjusted run (t-test p-values: LMC1: 6×10^{-4} , LMC4: 1.4×10^{-5} , LMC5: 3×10^{-3}), but only one component was mildly linked to sex in the execution of *MeDeCom* using the adjusted data matrix (LMC7, t-test p-value: 7.8×10^{-4} , Figure 4.4C). Although component 21 was associated with ethnicity in the ICA analysis, an equivalent association between any of the LMCs and donor ethnicity was not detectable. Surprisingly, neither age nor ethnicity were significantly linked to any component produced by either ICA or *MeDeCom*, in spite of the broad age range of 33 to 88 years. This indicated that age and ethnicity only marginally influence the DNA methylation patterns in this dataset.

Deconvolution Results

We employed the proposed protocol to the lung adenocarcinoma dataset from TCGA. Since no prior knowledge on the expected number of underlying cell types was available, the cross-validation procedure implemented in *MeDeCom* was employed for selecting the number of LMCs k . To that end, seven LMCs were selected, since the cross-validation error started to level off at this value of k (i.e., the elbow point of the cross-validation error curve in Figure 4.5), while the other metrics such as the objective value did not change. After fixing k to seven, we selected $\lambda=0.001$ (regularization parameter) as the point where the cross-validation error is still low, while the objective value substantially changes (Figure 4.5B). The DNA methylation values of the detected LMCs (entries of T) revealed LMC5 as particularly hypomethylated and LMC6 as highly methylated. In contrast, the remaining LMCs were rather intermediately methylated (Figure 4.5C).

As a next step, we associated the detected LMCs with biological properties to determine the biological implications of the components. As a first observation, LMC5 had substantially higher proportions in the healthy tissue samples in comparison to the tumor samples (Figure 4.6A). Thus, reference-free deconvolution was able to capture the inherent methylation signatures specific to cancerous and healthy tissues, which would typically be addressed using

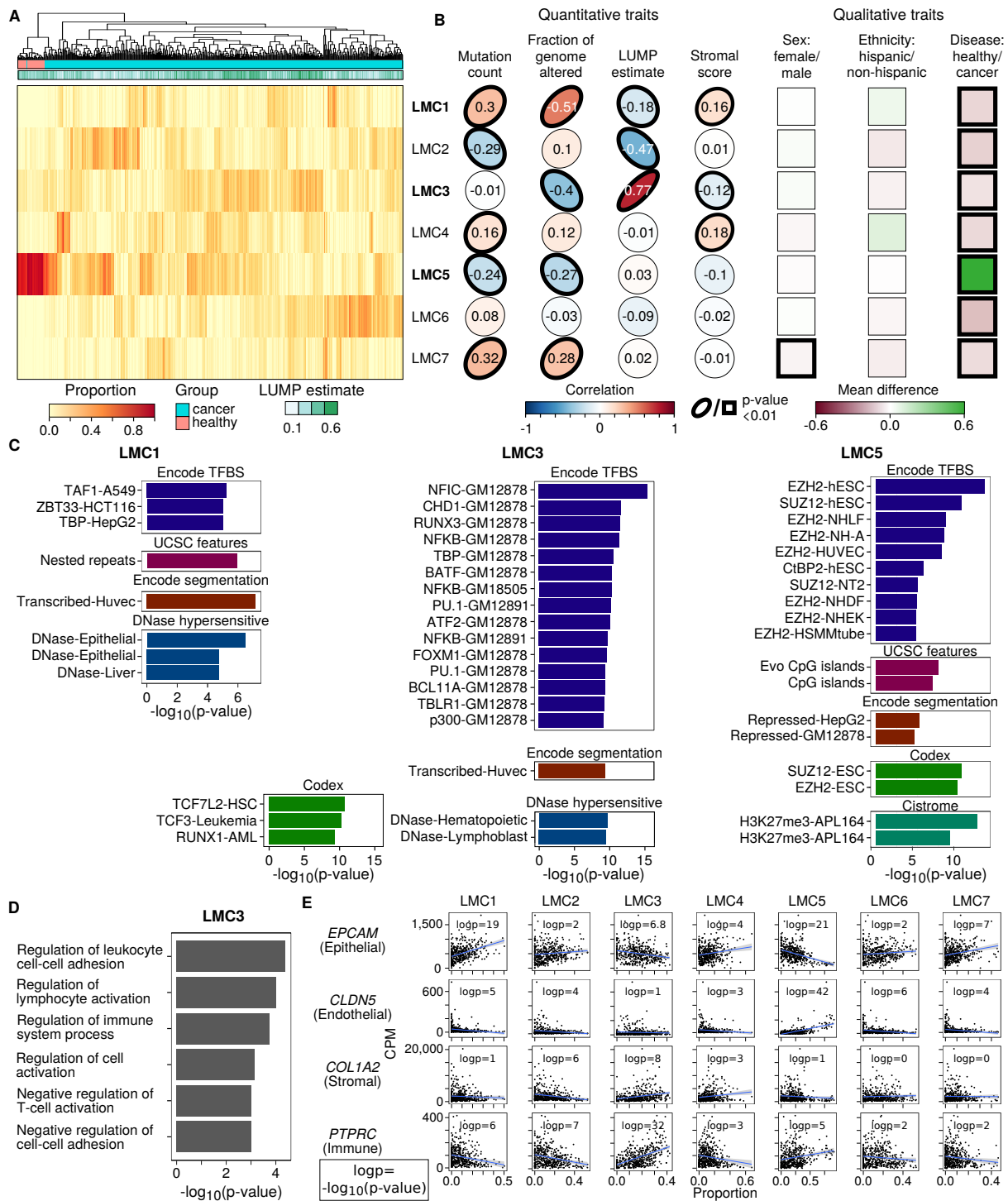


Figure 4.6: Interpreting *MeDeCom* results with *FactorViz* (Continued on next page).

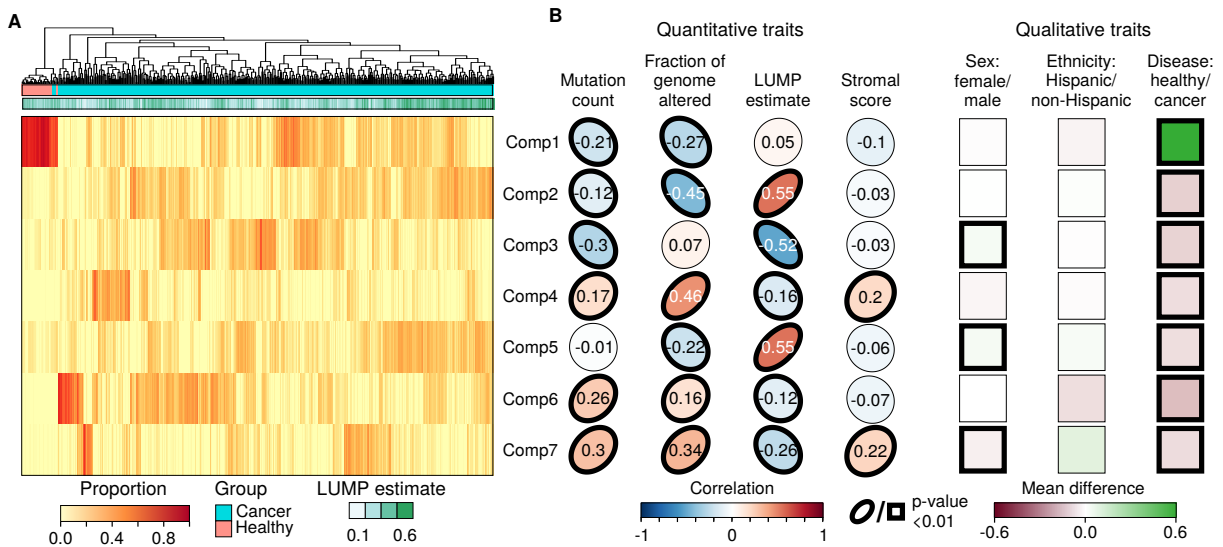
a differential analysis as described in Chapter 3. To associate LMCs with a subset of CpGs, we selected those sites that were specifically hypomethylated in an LMCs in comparison to the other LMCs. For LMC5, transcription factor binding sites of the core members of the Polycomb repressive complex 2 (PRC2), SUZ12 and EZH2, were overrepresented in the LMC5-specific

Figure 4.6: (Previous page) Interpreting *MeDeCom* results with *FactorViz*. **A:** Proportion heatmap of LMCs in the different samples ($k=7$ LMCs, $\lambda=0.001$). The samples were hierarchically clustered according to the Euclidean distance between the proportions using complete linkage. Samples are annotated using disease status and with the sample-specific LUMP estimate. **B:** Associations between selected phenotypic traits and LMC proportions. For quantitative traits, the Pearson correlations are shown as ellipses that are directed to the upper right for positive and to the lower right for negative correlations, respectively. For qualitative traits, the absolute difference of the proportions in the two groups (e.g., female vs. male) is shown. P-values below 0.01 are indicated by bold outlines. LOLA (**C**) and GO (**D**) enrichment analysis of the LMC-specific hypomethylated sites for LMCs 1, 3, and 5. No significant GO enrichment was found for LMCs 1 or 5. P-values have been adjusted for multiple testing with the Benjamini-Hochberg method [206]. Encode TFBS, transcription factor binding site ChIP-seq profiles from the Encode [168] project; Encode segmentation, chromatin state segmentation of Encode ChIP-seq profiles; Codex, ChIP-seq profiles from the Codex database [169]; Cistrome, ChIP-seq profiles from the Cistrome project [170]. **E:** Scatterplots between LMC proportions per sample and known marker gene expression of different lung cell types. Gene expression was quantified as counts per million (CPM), and the blue line represents the least squares regression line.

CpGs. PRC2 represses oncogenes and is often dysregulated in cancers, which is often in association with hypermethylation [253, 254, 255]. Additionally, LMC5 proportions in the tumor samples provided a tumor purity estimate, i.e., the proportions represented the degree of contamination by adjacent normal tissue. In summary, tumor-specific methylation signatures were captured with both *MeDeCom* (Figure 4.6) and *RefFreeCellMix* (Figure 4.7) without conducting differential analysis between two phenotypic groups. Notably, the ordering of the components is arbitrary and, for instance, *MeDeCom* LMC5 corresponds to *RefFreeCellMix* component 1.

LMC3 had highly variable proportions across the samples and was the main driver of the overall sample clustering without considering LMC5. LMC3 proportions were strongly correlated with the LUMP estimate (Figure 4.6B), which provides an estimate of the overall immune cell content of a sample based on CpGs that are particularly unmethylated in immune cells. Furthermore, LMC3-specific hypomethylated sites were enriched for leukocyte (B-lymphocyte) specific TFBS and immune response related GO terms (Figure 4.6C, D). In conclusion, LMC3 most likely represented tumor infiltrating immune cells. The extent of tumor infiltration might be relevant to associate cancer state to patient prognosis ([256], see also Section 4.2).

Lastly, we aimed at detecting cell-type specific profiles in the deconvolution results beyond cancerous and immune cell signals. Thus, we associated the detected LMC proportions with gene expression levels of known marker genes of lung tissue cell types. We selected *EPCAM* as an epithelial, *CLDN5* as an endothelial, *COL1A2* as a stromal, and *PTPRC* as an immune cell marker [257]. Gene expression data was retrieved from TCGA using the *edgeR* [258] and *TCGAbiolinks* [259] R-packages. LMC3 was correlated with *PTPRC* expression, while LMC1 was strongly associated with the expression of the epithelial marker *EPCAM* and LMC5 with the endothelial marker *CLDN5* (Figure 4.6E). This further indicated that reference-free deconvolution analysis is capable of detecting cell-type-specific DNA methylation signatures in cancer tissue samples.



4.1.4 Discussion

We developed a three-stage, computational protocol that facilitates reference-free deconvolution of complex DNA methylation data by providing systematic guidelines for dissecting DNA methylation data into its basic constituents. High-quality data is required as input to the deconvolution tool and the first stage of the pipeline presents the R-package *DecompPipeline* for data processing. Notably, the pipeline supports the published reference-free deconvolution tools *RefFreeCellMix*, *MeDeCom*, and *EDec*, but is readily extensible to new computational tools such as *TCA* [235] or *CONFINED* [236]. We found high concordance of results when applying *MeDeCom* or *RefFreeCellMix* as the deconvolution tool, respectively (cf. Figure 4.6, Figure 4.7). This further indicated that the choice of deconvolution tools is not as important as thorough data processing. Additionally, the pipeline might be useful for processing data for reference-based deconvolution, for example through the Houseman approach [224], *EpiDISH* [226], or *MethylCIBERSORT* [227]. The last stage of the protocol provides an interactive software application (*FactorViz*) for visualization, validation, and biological interpretation of deconvolution results. The execution of the protocol on lung adenocarcinoma data revealed that it is able to extract important biological features of solid tumors, including immune cell infiltration and tumor purity levels.

However, some limitations of the protocol should be mentioned. First, *MeDeCom* explores all user-specified combinations of the regularization parameter λ , the number of LMCs k , and several feature selection methods in a single execution. Thus, the number of basic deconvolution jobs can reach 1,000-10,000, which makes reference-free deconvolution a computationally

demanding task. A deconvolution analysis of a large dataset, including several hundreds to thousands of samples and hundreds of thousands to millions of selected CpGs, can take several days to finish even on larger machines. Notably, a standalone workstation was used for the analysis presented here, but the protocol provides functionality to execute the process on an HPC cluster environment. Additionally, biological interpretation and validation of the obtained LMC matrix requires user interaction and the results need to be checked for biological plausibility. To further automatize generating biological insights from deconvolution results, the graphical user interface *FactorViz* will be further improved to include new interpretation features and optimized user interaction. Lastly, accounting for confounding factors, especially for those that might have a strong influence on the methylome, can lead to a substantially altered overall DNA methylation data matrix. Currently, the proposed pipeline provides diagnostic plots, which require user interaction in order to determine whether and how the effect of a particular covariate is to be removed.

In the original publication of *MeDeCom*, the tool was validated on simulated data and *in-silico* cell-type mixtures. *MeDeCom* successfully identified neuronal and glial fractions in a brain frontal cortex dataset and detected additional LMCs associated with features of Alzheimer's disease [220]. We anticipate successful application of reference-free deconvolution in similar scenarios. Reference methylomes exist for blood-based studies [247], and reference-based methods tend to generate more reliable cell proportion estimates than reference-free deconvolution tools. However, in case of severely altered blood composition, e.g., due to an overproduction of rare cell types, the assumptions of reference-based deconvolution methods are violated and reference-free methods should be favored. For the analysis of blood samples and other similarly well-characterized tissues, *MeDeCom* can be applied in a semi-supervised fashion, i.e., the obtained LMCs can be compared to available reference profiles. This enables easy recovery of known signatures and allows for detection of additional unknown LMCs such as rare cell types.

For the future, we envision that single-cell resolution DNA methylation profiles will become increasingly available due to recent technological advances [260, 261]. Nevertheless, reference-free deconvolution of large-scale bulk tissue datasets will remain a necessary complement to single-cell DNA methylation profiling. This becomes especially relevant considering high costs, low sample throughput, and data sparsity of current single-cell applications [262]. More significantly, reference-free deconvolution can be used in combination with single-cell methylomes, e.g., single-cell profiling for several reference samples in controlled study settings followed by deconvolution of bulk methylomes from large patient cohorts. In such a setting, the single-cell resolution profiles can empower the interpretation of the LMCs, while deconvolution results can be used for interpretation and validation of single-cell profiles. Finally, since single-cell methylation maps are still especially hard to generate, deconvolution of more accessible bulk methylomes can be integrated with single-cell profiles of other data modalities. Most notably, single-cell transcriptomes (scRNA-seq) and chromatin accessibility maps (scATAC-seq, scNOME) can be integrated with deconvolution results from bulk experiments; an idea that is implemented in the recently published *EPISCORE* tool [263].

The presented protocol is not limited to Illumina microarray datasets, including Illumina EPIC and 450k data, but is readily extensible to bisulfite sequencing data (Supplementary Figure A.6). We expect the protocol to be of great benefit to all investigators performing DNA methylation analysis in complex and underexplored experimental systems, including bulk sam-

ples of highly heterogeneous tissues and tumors. We envision that, after further adaptations, the proposed protocol will also be applicable to datasets beyond DNA methylation, including transcriptomic data.

4.2 Reference-Free Deconvolution of DNA Methylation Data as a Prognostic Tool in Metastatic Melanoma

4.2.1 Lack of Predictors for Immune Checkpoint Inhibition Therapy Response in Melanoma

Throughout the last few years, the incidence of malignant melanoma in the United States substantially increased [264]. The stage of melanoma at the point of diagnosis critically influences patient prognosis, and stage IV (metastatic) melanoma patients generally have poor survival rates. Genetic mutations in the *BRAF* and *NRAS* genes are associated with enhancer tumor growth and therefore accelerate disease progression [265]. A substantial step forward in improving patient survival was the application of immune checkpoint inhibition (ICI) therapy, which improves overall patient survival [266]. ICI therapy modulates the immune system response to the tumor through blockage of specific immune checkpoints [267]. Immune checkpoints are regulatory pathways that prevent the immune system from attacking all cells without further considering the cell type. For the treatment of patients with metastatic melanoma, the combined blockage of *CTLA-4* and *PD-1* are in clinical use. However, a recurring issue with ICI therapy is resistance toward the treatment, and reliable predictors of ICI therapy resistance are missing. To exploit the potential of DNA methylation data as a potential predictor of ICI therapy success, we analyzed a cohort of stage IV melanoma patients under ICI treatment using the deconvolution protocol presented above.

4.2.2 Application of the Deconvolution Pipeline on Melanoma Data

Dataset Description

We analyzed 39 samples from skin metastases of patients with stage IV melanoma, which were treated with ICI therapy between 10/2010 and 01/2020. The following information was available for the samples: sex, age, *BRAF*/*NRAS* mutation status, and presence of brain metastases. Additionally, the patients were classified into ICI responders and non-responders according to criteria defining therapy success. The overall survival time, as well as the survival time from the first ICI treatment were recorded. DNA methylation data was generated using the Illumina EPIC microarray.

DNA Methylation Data Processing

DNA methylation data was obtained as IDAT files, which were used as input to *RnBeads*. Quality control was performed using the built-in control probes on the EPIC array, and the data showed high overall quality. Furthermore, CpGs were filtered according to detection p-values and annotated SNPs, sites on the sex chromosomes, and potentially cross-reactive sites were discarded from the analysis. Methylation data was normalized using the “dasen” method from the *wateRmelon* R-package. We used *RnBeads*’ exploratory and differential methylation module

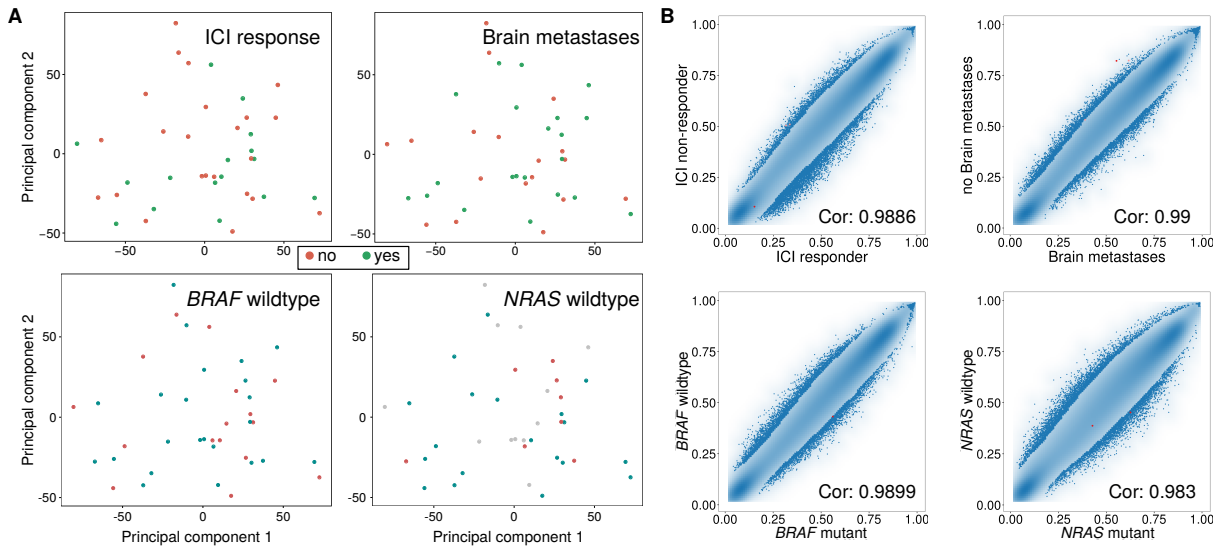


Figure 4.8: DNA methylation analysis of the melanoma cohort with respect to different sample annotations. **A:** PCA of genome-wide DNA methylation data at single-CpG resolution. Shown are the first two principal components along with the sample annotations ICI response, brain metastasis state, and *BRAF* and *NRAS* mutation states. **B:** Pairwise scatterplots of CpG-wise DNA methylation values averaged over all samples of the groups defined. The points in the low-point-density areas are drawn, while the points in high-point-density areas are visualized through kernel density estimation. Shown in red (only very few points) are the CpGs with a differential methylation p-value below 1×10^{-5} . The correlation is the Pearson correlation coefficient across all CpGs.

to determine associations of DNA methylation patterns with available sample annotations using PCA and differential methylation analysis. The data was further processed according to the deconvolution protocol presented above (see Section 4.1.2). We used ICA to adjust for age as a potential confounding factor and selected the 5,000 most variable CpGs as input to *MeDeCom*. Hierarchical clustering analysis (Euclidean distance, Ward's minimum variance method) on the LMC proportions across the samples yielded sample clusters that were investigated regarding their potential prognostic significance. To that end, we computed Kaplan-Meier survival curves using the *survival* R-package [249] and computed associated log-rank test p-values.

4.2.3 Prognostic Signature Identified Through Reference-Free Deconvolution

Associations between DNA Methylation and Sample Information

We found no associations of global DNA methylation patterns with available samples annotations in the PCA plots generated through *RnBeads* on the processed methylation data of the melanoma samples (Figure 4.8A). The sample annotations included classification into ICI responders and non-responders, the presence of brain metastases in the patients (yes vs. no), as well as the *BRAF* and *NRAS* mutation states (mutation vs. wildtype). This indicated that global DNA methylation variability was not driven by any of the available sample phenotypes, which was further confirmed by differential analysis (Figure 4.8B). Only a handful of CpGs reached the significance level of 1×10^{-5} in the differential analysis. After adjusting the p-values for multiple testing, none of the CpGs was identified as differentially methylated. Thus, we concluded that differential analysis is unable to detect reliable differences between ICI responders

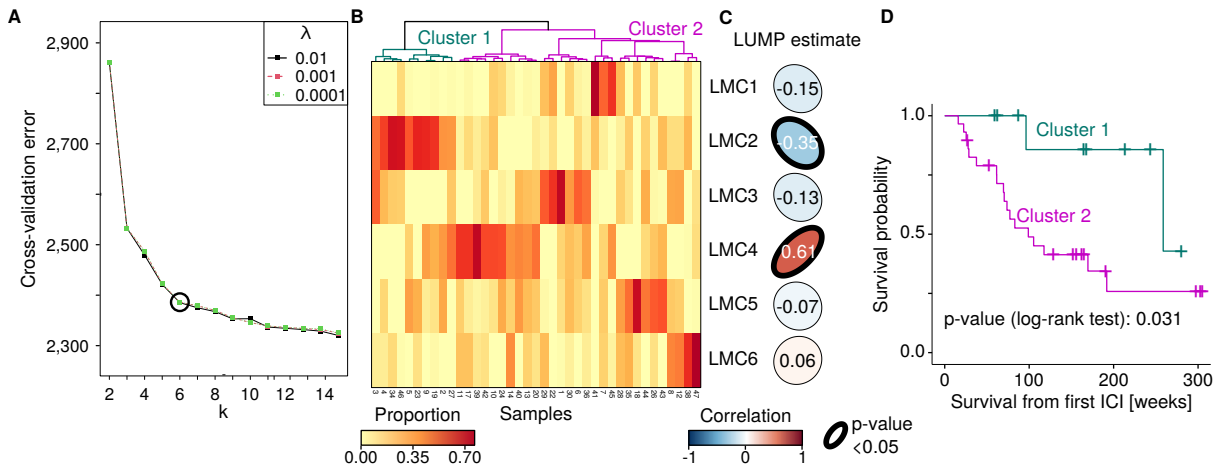


Figure 4.9: *MeDeCom* analysis of the melanoma cohort. **A:** Cross-validation error returned by *MeDeCom* plotted against the number of components k for different values of the regularization parameter λ . **B:** Heatmap of LMC proportions across the samples of the melanoma cohort. Samples were hierarchically clustered using Euclidean distance and Ward's minimum variance method. **C:** Pearson correlation between the LMC proportions and the sample-specific LUMP estimate. P-values according to a two-sided correlation test lower than 0.05 are indicated by a bold outline. **D:** Survival analysis using the hierarchical clustering information obtained on the proportions. The p-value was computed using the log-rank test.

and non-responders and we aimed to reveal those differences through deconvolution analysis.

LMC Clustering Informative about Patient Survival

To determine whether reference-free deconvolution analysis enables the detection of prognostic subgroups of samples, we applied the deconvolution protocol presented above to the melanoma cohort samples. We selected 6 LMCs, since the cross-validation error started to level off and selected 0.001 as the regularization parameter analogously (Figure 4.9A). The identified LMCs showed high proportions in different subgroups of the samples. For instance, LMC2 had high proportions in ten samples, which formed a distinct cluster (Figure 4.9B). By correlating the LMC proportions to available sample information such as the immune cell content estimated with the LUMP algorithm [149] we found a strong positive correlation of LMC4 with the immune cell content. In contrast, LMC2 exhibited a negative correlation with the LUMP estimate (Figure 4.9C). We used hierarchical clustering on the proportions to obtain subgroups of samples. Two clusters were identified and we compared the ten samples with substantially higher LMC2 proportions (Cluster 1) with the remaining samples (Cluster 2) in a survival analysis. More specifically, we used the survival time from the first ICI treatment as output variable in the log-rank test and found a significantly better survival for the ten samples in Cluster 1 (Figure 4.9D). Notably, Cluster 1 included the samples with higher LMC2 proportions, which had negative correlation with the estimated immune cell content. This means that higher immune cell infiltration into the tumor correlated with shorter survival time from the first ICI treatment in this cohort.

4.2.4 Discussion

There is need of a prognostic biomarker for predicting the success of ICI therapy applied to patients with metastatic melanoma. In this project, we exploited the potential of DNA methylation data as a predictor of ICI treatment success. First, we found no indications of a DNA methylation difference associated with ICI therapy resistance using *RnBeads*. Thus, we used the reference-free deconvolution protocol on the 39 melanoma patients and found a subgroup of patients with high proportions of LMC2, which showed a significantly better survival than the remaining samples. LMC2 proportions negatively correlated with the predicted immune cell content of the samples indicating that low immune cell infiltration into the tumor is beneficial for patient survival. We would like to point out that the findings require validation using independent datasets, since we investigated a rather small cohort of patients. We were unable to validate the findings on TCGA data, since only few stage IV tumors are present in the TCGA cohort and information about ICI therapy is absent. Additionally, we could not find an enrichment toward a common regulatory role or a biological pathway for the CpGs that are specific to LMC2. After validation of the findings, the characteristic properties of LMC2 could be more thoroughly investigated to construct a DNA methylation-based predictor of ICI therapy success.

DNA Methylation Heterogeneity Within Samples

In this chapter, I discuss the third level of DNA methylation heterogeneity – heterogeneity within biological samples or within-sample heterogeneity (WSH) – as an important layer of information in DNA methylation data that is commonly neglected. I will focus on bisulfite sequencing data in this chapter, since WSH can be reliably estimated from bisulfite sequencing reads. Together with Markus List and Fabian Müller, who mainly supervised this work, I set out to systematically investigate genome-wide scores for quantifying WSH. Within this project, which is a modified version of the work published as Scherer et al. [268] in Nucleic Acids Research (2020), I developed a novel mathematical score for quantifying WSH at single-CpG resolution. Using simulated and experimental data, I observed that the new score – $qFDRP$ – was less affected by technical biases and reliably quantified WSH originating from different sources of WSH in our simulation experiments. While this chapter mainly focuses on the mathematical aspects of WSH scores and aims to use computational approaches for quantifying WSH, I also present a potential biological application of the score as a reliable estimator of tumor purity.

5.1 Quantitative Comparison of Within-Sample Heterogeneity Scores for DNA Methylation Data

5.1.1 Within-Sample Heterogeneity in DNA Methylation Data

In the previous chapters, the main focus was on data generated using the Illumina BeadArrays. While these microarrays allow for the generation of large-scale datasets, bisulfite sequencing approaches, including RRBS and WGBS, provide information beyond the methylation state of a single CpG. Each of the sequencing reads assayed using RRBS/WGBS comprises information about the relationship between the methylation states of multiple CpGs [269, 270] and is accordingly informative about co-methylation patterns. Since the methylation information per sequencing read per CpG is obtained from a single molecule, only two potential states exist: methylated or unmethylated. However, when comprehensively investigating the average DNA methylation state across the sequencing reads for all CpGs in various tissues, such as breast, blood, brain, and embryonic stem cells, Elliott et al. [271] found that about 2% of 26.9 million CpGs showed intermediate DNA methylation values (higher than zero, but lower than one). These intermediate values mainly arise from within-sample heterogeneity (WSH), which can originate from, among others, cell-type heterogeneity, cellular contamination, allele-specific DNA methylation (ASM), and DNA methylation erosion (Figure 5.1).

To put this in other words, the observed DNA methylation level (computed with Equation 2.2)

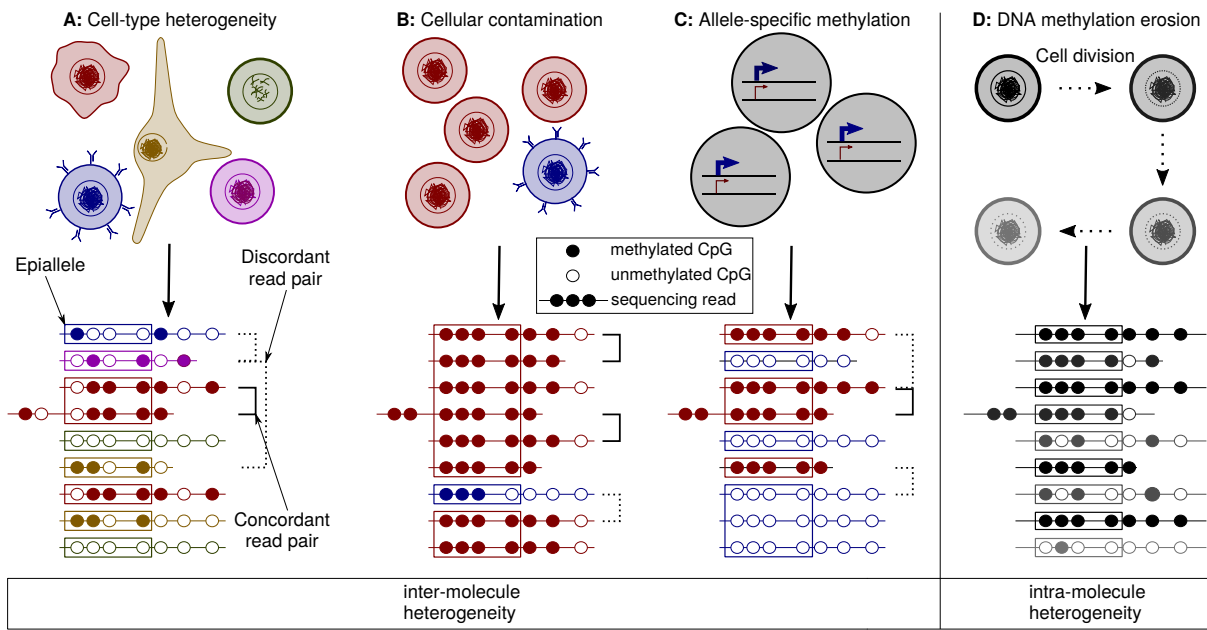


Figure 5.1: Sources of WSH and their manifestation in bisulfite sequencing reads. **A:** Cell-type heterogeneity: a sample comprises different cell types with unique DNA methylation patterns. **B:** Cellular contamination: cell sorting during sample preparation does not yield a pure population of cells of interest. **C:** ASM: two alleles differ in their DNA methylation states and each of the states is reflected in the sequencing reads. **D:** DNA methylation erosion: cells lose DNA methylation in multiple rounds of cell division in a stochastic process.

represents the average profile of a variety of distinct DNA methylation states in a population of different cellular states/cell types [272, 248]. WSH is manifested in heterogeneous DNA methylation patterns observed in the sequencing reads at a given genomic location. To partially account for these heterogeneous patterns, bulk tissues can be physically separated into cell types or tissue regions [80, 273] using, e.g., FACS. Additionally, *in-silico* approaches including reference-free deconvolution tools (see Chapter 4) can be used or bisulfite sequencing reads can be clustered according to their putative cell-of-origin [103]. Here, we aim at quantifying these heterogeneous patterns and use them as a feature rather than as a confounding factor and determine regions of elevated WSH in the genome or regions with differential heterogeneity between two phenotypic groups.

While single-cell bisulfite sequencing has the potential of overcoming the issues introduced by WSH, large case-and-control studies are currently not feasible due to high costs and technical challenges [262]. Furthermore, a number of technological challenges have to be solved before single-cell datasets reach the quality of published, large-scale bulk datasets, such as the ones generated by international epigenomic consortia. On the other hand, local deep amplicon sequencing enables estimating the true probability distribution of DNA methylation patterns in a biological sample [274, 275]. Local deep amplicon sequencing generates high-resolution DNA methylation profiles at single genomic loci, but does not afford genome-wide coverage [97]. WSH scores quantify genome-wide WSH and are in the focus for this chapter.

Landau et al. [276] introduced the *Proportion of Discordant Reads* (PDR) for quantifying locally disordered DNA methylation patterns as one of the first genome-wide WSH scores. PDR exploits the correlation structure between DNA methylation states of neighboring CpGs on the

same sequencing read. Guo et al. [277] proposed *Methylation Haplotype Load* (MHL), which quantifies fully methylated substrings in the reads to construct methylation haplotype blocks. *Epipolymorphism* [278] and *(Methylation) Entropy* [279] define an epiallele as a configuration of DNA methylation states within a window of 4 CpGs. Using epiallele frequencies in the reads, the variance at a genomic locus is quantified using Shannon and Tsallis Entropy, respectively. We developed two new scores – the *Fraction of Discordant Read Pairs* (FDRP) and *quantitative FDRP* (qFDRP) – that quantify WSH at single-CpG resolution using pairwise distances between reads.

WSH scores describe the variance in sequencing patterns and thus describe another aspect of DNA methylation in comparison to the DNA methylation level. Previous studies showed that PDR and Entropy are associated with gene expression [276] and transcriptional heterogeneity [82]. Additionally, PDR correlated with important clinical parameters including tumor size, progress-free survival, and tumor location [83, 154]. PDR was surprisingly lower in chronic lymphocytic leukemia (CLL) patients than in healthy controls and showed a variety of correlations with gene expression states [82].

To facilitate choosing the most appropriate score for an analysis, a systematic review of WSH scores is needed. Therefore, we evaluated PDR, MHL, Epipolymorphism, Entropy, FDRP, and qFDRP in the context of simulated and publicly available bisulfite sequencing data. More specifically, three criteria were used to evaluate performance: First, we tested if WSH scores capture different sources of heterogeneity in simulated data (Figure 5.1). Second, we assessed robustness of WSH scores with respect to CpG density and technical biases, such as sequencing coverage and read length. Third, we investigated the biological implications of elevated WSH. To that end, we checked whether WSH scores reveal novel regulatory regions in particular those that are not apparent on the average DNA methylation level and we associated WSH scores with tumor purity estimates.

5.1.2 Description of WSH Scores and Data Simulation

Definitions of WSH Scores

General Definitions As a first step, we introduce a unified mathematical representation of the WSH scores (see also Table 5.1) investigated, which requires some general definitions:

$$\begin{aligned} r &= \text{set of CpG positions (representing a read)} \\ R_c &= \text{set of all reads } r \text{ covering } c \\ x_{i,r} \in \{0, 1\} &= \text{methylation state of CpG position } i \text{ in read } r \end{aligned}$$

Here, c is the CpG of interest. Furthermore, 0 represents an unmethylated CpG and 1 a methylated CpG.

PDR The Proportion of Discordant Reads (PDR [276]) quantifies locally disordered DNA methylation states of CpGs on the same sequencing read. Reads are classified as concordant, if all CpGs on the read have the same methylation state and as discordant otherwise. Following

Table 5.1: Characteristics of different WSH scores targeted at quantifying inter-molecule and intra-molecule heterogeneity.

Type	inter-molecule				intra-molecule	
Biological assumption	each epiallele originates from single allele in single cell	each read originates from single allele in single cell			local concordance of CpG methylation states	
Concept	computing frequencies	epiallele	pairwise concordance of reads		agreement between neighboring CpGs	
Similarity calculation	Tsallis entropy of epiallele frequency	Shannon entropy of epiallele frequency	pairwise discordance of reads	pairwise distance of reads	all CpGs on read either methylated or unmethylated	number of consecutively methylated substrings
	Epipoly	Entropy	FDRP	qFDRP	PDR	MHL

this classification, PDR is defined for a CpG c as:

$$\text{PDR}(c) = \frac{\sum_{r \in R_c} I(\exists i, j \in r \text{ s.t. } x_{i,r} \neq x_{j,r})}{|R_c|}$$

Note that the indicator function I is 1, if CpGs at any two positions in the read (i, j) have different DNA methylation states (read is discordant). As a requirement stated in the original publication and as a threshold used throughout this work, reads are only included in the read set R_c if they contain at least four CpG sites.

MHL The Methylation Haplotype Load (MHL, [277]) determines fully methylated substrings of differing lengths in each of the reads and computes the proportion of methylated substrings over all possible substrings. It is defined for a given CpG c as:

$$\text{MHL}(c) = \frac{\sum_{l=0}^L (l+1) \frac{\sum_{r \in R_c} \sum_{i=1}^{|r|-l} I(x_{i,r} = 1 \wedge \dots \wedge x_{i+l,r} = 1)}{\sum_{r \in R_c} |r| - l}}{\sum_{l=0}^L (l+1)}$$

$|r|$ = number of CpGs in read r

$l+1$ = number of consecutive CpGs with identical methylation states

$L = \max_{r \in R_c} (|r|) - 1$

Epipolymorphism and Methylation Entropy Epipolymorphism [278] and Methylation Entropy [279] quantify the variance of *epialleles*, which are configurations of methylation states

in windows comprising four CpGs (2^4 potential configurations). The frequency of each of the epialleles is calculated from the sequencing reads and Epipolymorphism for window w is computed as:

$$\text{Epipolymorphism}(w) = 1 - \sum_{k=1}^{16} p_k^2$$

$$p_k = \frac{\sum_{r \in R_w} I(\forall i \in c_k : x_{i,c_k} = x_{i,r})}{|R_w|}$$

$c_k \in \{(0, 0, 0, 0), (0, 0, 0, 1), \dots, (1, 1, 1, 1)\}$ (epiallele)

R_w = set of all reads r containing all four CpGs in w

$x_{i,c_k} \in \{0, 1\}$ = methylation state of CpG i in epiallele c_k

w = window of four consecutive CpGs

Similarly, methylation entropy is calculated as:

$$\text{Entropy}(w) = -\frac{1}{4} \sum_{k=1}^{16} p_k \times \log_2 p_k$$

Epipolymorphism and Entropy compute the entropy of DNA methylation patterns observed in windows of four consecutive CpGs (w) across the sequencing reads.

FDRP and qFDRP As a novel score, we introduce the Fraction of Discordant Read Pairs (FDRP), which captures within-sample DNA methylation heterogeneity at single CpG resolution. FDRP is defined as:

$$\text{FDRP}(c) = \frac{\sum_{r_s \in R_c} \sum_{r_t \in R_c, t > s} I(\exists i \in \{r_s \cap r_t\} \text{ s.t. } x_{i,r_s} \neq x_{i,r_t})}{\binom{|R_c|}{2}}$$

r_s, r_t = sets of CpG positions (representing reads)

$s, t \in [1, |R_c|]$ = indices of reads

FDRP takes all read pairs into account that cover the sequence position of interest (c). A read pair is classified as discordant, if there is a CpG with different DNA methylation states in the two reads. FDRP is the fraction of discordant read pairs among all read pairs.

The quantitative FDRP (qFDRP) is derived from FDRP as follows:

$$\text{qFDRP}(c) = \frac{\sum_{r_s \in R_c} \sum_{r_t \in R_c, t > s} \frac{\sum_{i \in \{r_s \cap r_t\}} I(x_{i,r_s} \neq x_{i,r_t})}{|\{r_s \cap r_t\}|}}{\binom{|R_c|}{2}}$$

r_s, r_t = sets of CpG positions (representing reads)

$s, t \in [1, |R_c|]$ = indices of reads

qFDRP computes the Hamming distance between the methylation states of all the read pairs, which replaces the definition of discordance used by FDRP. For FDRP and qFDRP, the number of pairwise comparisons increases quadratically with read coverage, which is addressed for using a subsampling strategy (see Implementation).

Implementation

FDRP and qFDRP have been implemented in the R-package *WSHPackage*, which is available from GitHub¹. The following parameter settings are the default values implemented in the package and also the ones employed in the presented analysis. To avoid the issues of the quadratic growth, we randomly subsampled 40 reads at genomic regions with coverage higher than 40 reads. Furthermore, we discarded read pairs overlapping by fewer than 35 bp to focus only on read pairs with shared information. By using a fixed-sized window (50 bp) around the CpG site of interest, the scores are also applicable to datasets with different read lengths. *RnBeads* data structures were used for storing DNA methylation, coverage and sample information. To focus on a set of reliably covered regions, we only used CpG sites with read coverage at least 10 in the experimental data to compute FDRP and qFDRP.

Epiallele frequencies were computed using the *methclone* software (version 0.1 [280]), and the output was used to calculate Epipolymorphism and Entropy using custom R scripts. We used the following parameters for *methclone*: `methylation difference: 0`, `distance cutoff: 50 bp`, and `coverage threshold: 10`. PDR was calculated from the aligned sequencing reads using custom R scripts, and MHL was quantified using the Perl scripts provided by the original authors [277].

For scheduling compute jobs across the nodes of a HPC cluster, we used the python packages *peppy*² and *pypiper*³. All the plotting scripts and pipelines are available from GitHub⁴, which are recommended for analyzing large datasets comprising tens to several hundreds of samples. The R-package (*WSHPackage*) available from GitHub implements FDRP, qFDRP, PDR, MHL, Epipolymorphism, and Entropy and comprises an extensive vignette and manual. As input, the package requires aligned reads in BAM format and genomic annotation for the CpGs of interest through an *RnBiseqSet* or *GRanges* object [207].

Simulation of Bisulfite Sequencing Reads and Evaluation of WSH Scores

Simulation Setup We simulated bisulfite sequencing reads from the human reference genome version ‘hg38’ (chromosomes 22, X, and Y excluded) using the *Sherman* tool⁵. *Sherman* defines the methylation probability of each CpG through the `--CG_conversion` parameter, which implicitly controls sample heterogeneity since a methylation probability of 50% results in completely random, i.e., very heterogeneous patterns. We used different parameter settings for the different scenarios. Simulated reads were aligned to the reference genome ‘hg38’ with *bismark* (version 0.13.0 [129]) to create BAM files. DNA methylation scores were extracted and processed

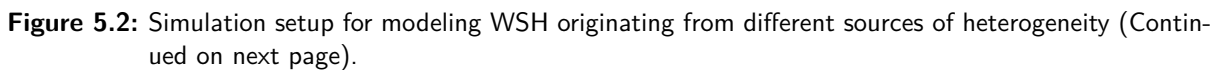
¹<https://github.com/MPIIComputationalEpigenetics/WSHPackage>

²<https://github.com/pepkit/peppy>

³<https://github.com/databio/pypiper>

⁴<https://github.com/MPIIComputationalEpigenetics/WSHScripts>

⁵<https://www.bioinformatics.babraham.ac.uk/projects/sherman/>



Simulated Heterogeneity For each source of WSH investigated here (Figure 5.1), bisulfite sequencing reads were simulated in 1,000 randomly selected genomic regions of length 50 kb (a simulated dataset) using *Sherman* (read length: 50 bp, error level: 1%). For each simulated dataset, we created different subpopulations of reads representing different cellular states (such as cell types, see Figure 5.2). These subpopulations were subsequently merged to generate simulated samples for each dataset and scenario separately. For each subpopulation of reads in each of the simulated datasets, we selected the same background methylation state (fully

Figure 5.2: (Previous page) Simulation setup for modeling WSH originating from different sources of heterogeneity. **A:** We merged between 2 and 10 simulated cell types *in-silico* per dataset, each with a randomly selected DNA methylation state within a randomly selected sub-region. **B:** A pure cell population is 'contaminated' with 10-50% of another cell type. **C:** Two cell types were generated, with one changing its DNA methylation state in a random sub-region. **D:** In a randomly selected sub-region, the fully methylated background state of all CpGs on a given read is de-methylated with the same probability α . Those 'eroding cells' are replicated γ times to represent the stochasticity of selecting cells for sequencing from a population. **E:** No heterogeneity was introduced. In a single cell type, the DNA methylation state changes from fully methylated to unmethylated or *vice versa*. THR, truly heterogeneous region

methylated/unmethylated, with error level 1%) as baseline. Within each of the subpopulations individually, we introduced the opposite methylation state in a subset of CpGs within a randomly selected subregion. We define the truly heterogeneous region (THR) per simulated dataset as the union of all the subregions with the opposite methylation state across all the read subpopulations.

To assess the potential of each score to quantify heterogeneity, we performed a t-test comparing the different scores at each CpG in the background to the CpG-wise scores within the THR. Additionally, we simulated negative cases in which no change in DNA methylation state for the subregion was introduced. In total, we created 1,000 simulated regions, which comprise a THR in about half of the regions for each simulation scenario separately. We used these definitions to determine the numbers of true positives, true negatives, false positives, and false negatives, as well as resulting ROC curves for each score and scenario.

To model cell-type heterogeneity, we merged a random number (between 2 and 10) of simulated cell types (read subpopulations). Within each of these cell types, a random subregion was introduced showing the inverse DNA methylation state in comparison to the baseline. We defined the THR for each of the 1,000 simulated regions as the maximum segment at which any of the cell types changes from baseline DNA methylation level. Then, we computed the average WSH score for each of the regions individually and correlated this quantity with the number of cell types in this particular region to determine if the scores quantify the degree of heterogeneity in a sample. In the cellular contamination scenario, only two distinct cell types were mixed at a random proportion between 1.0 and 0.5 (Figure 5.2). Analogously, ASM was simulated using 0.5 as fixed proportions for the two cell types.

For simulating DNA methylation erosion, we first generated fully methylated regions for one cell type. For each simulated dataset, we randomly selected a subregion in which we demethylate CpGs with probability α . Since the eroded segments are passed on to all daughter cells in cell replication, we sampled from these demethylated reads between 2 and 10 times (parameter γ), which also reflects the random selection of fragments for sequencing.

Borders of active regulatory elements are often marked by a sharp increase/decrease in the DNA methylation state, and we refer to this scenario as a methylation switching domain (MSD). In this scenario, we modeled a single cell type with a either fully methylated or unmethylated baseline DNA methylation level. Within a randomly selected subregion, the cell type changes its DNA methylation state to the inverse state.

Technical Biases We simulated additional datasets to address three potential technical biases arising from bisulfite sequencing that might affect the WSH scores: read coverage, read length, and sequencing errors. Similar to the scenarios described above, we randomly selected 1,000 regions of size 50 kb and simulated reads using *Sherman*. In order to investigate the effect of technical biases without considering the different sources of WSH described above, we computed 62.5% as the average DNA methylation level in the blood cohort (see subsequent section). In this dataset, 42.4% and 26.4% of all sites had methylation levels of at least 95% and of at most 5%, respectively. Thus, we used an overall methylation probability of 62.5% and set the `--CG_conversion` parameter to 95 and 5 for the methylated and unmethylated states, respectively. For modeling datasets with different read coverages, we created another 1,000 simulated regions. For each of the regions individually, we selected between 5,000 and 50,000 reads (step size 5,000) representing coverage between roughly 5- and 50-fold. Similarly, we employed different read lengths (`--length` parameter) increasing from 40 to 150 bp (step size 10). We changed the number of reads generated according to the length parameter in order to keep CpG-wise read coverage constant across the different regions. Lastly, we introduced different sequencing error levels (1 to 10 percent, step size 1%) using the `--error_rate` parameter in *Sherman*. *Sherman* employs an exponential decay model for each nucleotide with lower error probability for the 5' than for the 3' end of the reads (see also *Sherman* manual⁶).

Experimental Data

We analyzed RRBS data comprising 239 whole blood samples to validate findings on simulated data. The dataset comprises whole blood samples of healthy individuals, and focused on human longevity. Therefore, the dataset spans a range of 20-103 years of age for the individuals within the cohort. The dataset is accessible through the PopGen Biobank⁷ [165]. 5,606,227 CpGs were covered at average read depth 7.5. We also used this blood cohort to estimate parameters for the simulation experiments.

To illuminate the WSH scores' abilities to quantify WSH irrespective of the sequencing technology used, we collected hepatic WGBS samples from DEEP (European Genome-phenome Archive, EGA accession: EGAD00001002527). More specifically, we artificially mixed WGBS data from a hepatic cancer cell line (HepaRG) and a primary hepatocyte sample to generate a heterogeneous WGBS sample. In order to contrast WSH scores in heterogeneous and homogeneous samples, we also generated a homogeneous WGBS sample. To achieve that, we mixed two primary hepatocyte samples, which are expected to be more similar to one another than to the HepaRG sample. Library preparation and sequencing was conducted by Gilles Gasparoni and primary data processing was performed according to the DEEP WGBS process documentation⁸ by Karl Nordström from the Department of Genetics/Epigenetics. The final dataset covered 23,290,153 sites at average read depth 24.1.

Lastly, a RRBS dataset comprising samples obtained from Ewing sarcoma patients (GEO accession: GSE88826 [154], see also Chapter 3) was used to illustrate the applicability of WSH scores in a disease context. This dataset comprised 188 samples with 140 Ewing tissue samples, 16 Ewing cell lines (Ewing CL), 21 mesenchymal stem cells extracted from healthy donors (MSCs), and 11 MSCs extracted from Ewing sarcoma patients (eMSCs) and covers 2,217,786

⁶<https://www.bioinformatics.babraham.ac.uk/projects/sherman>

⁷<https://www.uksh.de/p2n/>

⁸<https://github.com/molgen.mpg.de/DEEP/comp-metadata>

sites at average read depth 14.7. We used *TrimGalore!*⁹ for trimming and aligned the sequencing reads to reference genome version 'hg38' using *bsmap* [128].

Quantification of WSH Scores, Tumor Purity, and Differential Heterogeneity

In a first processing step, we converted sample-specific WSH scores into a data matrix of dimension CpG sites \times samples. To focus on potentially regulatory regions of the genome, single-CpG WSH scores were further aggregated across samples or across putative functional elements according to the Ensembl Regulatory Build [8]. We excluded 11 formalin-fixed and paraffin-embedded (FFPE) samples from the Ewing tissue group, since they showed lower quality in the original publication [154].

Tumor purity levels, or the level of immune infiltration into the tumor microenvironment, are important information for therapy selection (cf. Section 4.2). These scores are typically obtained from histopathological investigation of the sample or from genetic data [149, 282, 283]. However, in the absence of such information, WSH scores could serve as estimates of tumor purity since they quantify heterogeneity. In the Ewing cohort dataset, tumor purity levels were estimated for 81 samples using genetic data based on loss of heterozygosity, copy number change, and the mutated allele fraction [154] with the method described in Chen et al. [284]. To estimate tumor purity levels from WSH scores for the remaining samples, we trained an elastic net regression model using the *glmnet* R-package [136]. Elastic net regression accounts for the high dimensionality of the problem using regularized linear regression (see Section 2.6 for details). We used ten different initializations of 10-fold nested cross-validation to select the elastic net hyperparameters α and λ simultaneously. While α determines the weight of either the ridge regression penalty ($\alpha=0$) or the Lasso penalty ($\alpha=1$), λ determines the weight of the regularization in comparison to the objective value (Section 2.6). Model performance was only marginally affected by selecting α and we thus selected $\alpha=1$, since the Lasso returns simpler models (i.e., comprising less features) than ridge regression. Then, we selected those sites that consistently had non-zero coefficients in five or more folds when we executed ten different initialization of ten-fold cross-validation of the Lasso. To validate that we did not find an association by chance, we randomized sample labels and re-ran the regression model. Subsequently, we used the selected sites for each of the WSH scores individually and conducted another ten-fold cross-validation using an unregularized linear least squares model to estimate overall performance of the proposed method. We visualized the WSH scores of the selected sites using heatmaps.

We used the logit-transformation of the scores (M-values) to determine differentially methylated and differentially heterogeneous regions between two groups using linear models implemented in the *limma* R-package [246]. The analysis was performed both on the single-CpG level and after aggregation across putative regulatory elements. We adjusted the resulting p-values for multiple testing using the Benjamini-Hochberg method [206] and employed a FDR threshold of 0.01 to determine differentially heterogeneous sites. These sites were used as input to enrichment analysis with *GOstats* [250] and *LOLA* [152].

⁹http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

5.1.3 Application of WSH Scores on Simulated and Experimental Data

Conceptual Comparison of WSH Scores

Before computing the WSH scores on simulated data, we compared the scores based on their construction and biological motivation. WSH scores can be conceptually divided into two classes: *intra-molecule* scores exploit the (dis)agreement of CpG-wise DNA methylation states on the same sequencing read, while *inter-molecule* scores quantify the variability of DNA methylation patterns observed in the reads at a genomic locus. Intra-molecule scores are mainly motivated from DNA methylation erosion and inter-molecule scores were created to capture cell-type heterogeneity (Table 5.1).








Intra-Molecule Scores: DNA Methylation Erosion DNA methylation patterns in cancer are intrinsically heterogeneous and PDR has originally been developed to describe locally disordered DNA methylation states. For any given CpG, PDR is computed as the fraction of reads that are discordant. Thus, PDR is maximal (1), if all reads that contain a specific CpG comprise both methylated and unmethylated CpGs (Table 5.2). Accordingly, PDR is minimal (0) if all reads are consistently methylated/unmethylated. MHL defines DNA methylation haplotypes based on the CpG methylation states on a read. It is maximal (1) if all reads are fully methylated and minimal (0) if they are completely unmethylated. MHL does not increase linearly with the number of methylated CpGs but rather quantifies stretches of adjacently methylated CpGs.

Inter-Molecule Scores: Cell-Type Heterogeneity We differentiate between two sub-classes of inter-molecule WSH scores; Epipolymorphism and Entropy utilize *epialleles* as combinations of CpG methylation patterns within four-CpG windows, while FDRP and qFDRP conduct pairwise comparisons of sequencing reads. Detecting epiallele frequencies in the reads is an important prerequisite for computing Epipolymorphism and Entropy. For a DNA methylation pattern of four CpGs ($2^4=16$ combinations in total), the frequency of each of these epialleles in the reads is determined. Epipolymorphism quantifies Tsallis entropy on the frequencies and is maximal ($1 - \frac{1}{16}=0.9375$), if all 16 epialleles occur at the same frequency. In accordance, Entropy utilizes Shannon entropy and is minimal if a single pattern (e.g., one cell type) is present in the reads. Both scores rely on spatial proximity of CpGs, since multiple (four) CpGs have to be present in the reads.

FDRP and qFDRP compare pairs of reads/patterns. While FDRP classifies each pair of reads into either *concordant* (if all DNA methylation states match) or *discordant* (if one or more DNA methylation states differ), qFDRP computes the Hamming distance of DNA methylation states on the two reads. FDRP is maximal (1) if no two reads reflect the same DNA methylation configuration and qFDRP/FDRP are lowest (0) if all reads agree. qFDRP is maximal (1), if none of the methylation states in any of the read pairs agree, which can only occur with two distinct reads. Notably, FDRP is an upper bound for qFDRP and qFDRP can be considered a soft version of FDRP.

Comparison Between Intra- and Inter-Molecule WSH Scores An important property of WSH scores is that they identify differences in DNA methylation patterns beyond those captured by the average DNA methylation level. For instance, the DNA methylation level is the

Table 5.2: Comparison of average DNA methylation level and WSH scores for different sequencing read configurations. n.d.=not defined

Reads							
DNAm level	0	0.5	0.5	0.5	0/1	1	0.667
FDRP	0	0.6	0.6	1	0	0	0.733
qFDRP	0	0.6	0.6	0.6	0	0	0.555
PDR	0	0	1	1	1	0	1
MHL	0	0.5	0.117	0.083	0.358	1	0.134
Epipoly	0	0.5	0.5	0.83	0	0	<n.d.>
Entropy	0	0.25	0.25	0.65	0	0	<n.d.>

same for read configurations 2-4 in Table 5.2, while the read patterns are largely distinct; a property of the DNA methylome for which WSH scores account. MHL is conceptually different from the other scores, since it quantifies stretches of adjacently methylated CpGs rather than heterogeneity in methylation patterns. However, due to its construction, it shares properties of the DNA methylation level in fully methylated and unmethylated regions (Table 5.2). Since large fractions of the human genomes are either fully methylated or unmethylated, MHL and the DNA methylation level are expected to be identical for the majority of the genome. Although PDR is motivated from DNA methylation erosion and FDRP/qFDRP rather from cell-type heterogeneity, they share certain characteristics. The three scores (qFDRP, FDRP, PDR) conceptualize the concordance either between neighboring CpGs on the same read (PDR) or for the same CpG on different reads (FDRP/qFDRP). Nevertheless, PDR and FDRP/qFDRP are similarly elevated in disordered situations (cf. Configuration four in Table 5.2). On the other hand, there are read configurations in which either inter-molecule scores or PDR are high, while the other scores are low (Table 5.2, Table 5.3). If a pair of reads contains many overlapping CpGs, the probability of detecting a single difference increases and results in a discordant read pair according to the definition used by FDRP. This was the original motivation for the development of qFDRP, which is less susceptible to distinct DNA methylation states occurring at single CpGs at very low frequencies. Epipolymorphism and Entropy share the definition of epialleles and use entropy to measure the variance and thus behave similarly across different read configurations.

Cell-Type Heterogeneity, Cellular Contamination, and ASM Captured by Inter-molecule WSH Scores in Simulation Experiments

Cell-Type Heterogeneity, Cellular Contamination, and ASM Cell-type heterogeneity critically influences the analysis of bulk tissue samples using bisulfite sequencing and is typically considered the major confounding factor in any epigenomic study. In our first simulation scenario, we *in-silico* merged 2-10 simulated cell types (Figure 5.2). The performances of WSH scores were evaluated based on truly heterogeneous regions (THRs), which we introduced as regions with a substantially elevated number of cell-type-specific patterns and thus with elevated WSH. We found that the DNA methylation level, Epipolymorphism, Entropy, FDRP, and qFDRP correctly identified the THRs in which the cell types exhibited distinct DNA methy-

Table 5.3: Examples of read configurations and resulting WSH scores.

Reads	FDRP	qFDRP	PDR	MHL	Epipoly	Entropy
	$\frac{3}{3} = 1$	$\frac{\frac{1}{4} + 1 + \frac{3}{4}}{3} = \frac{2}{3}$	$\frac{1}{3}$	$\frac{1 \times \frac{7}{12} + 2 \times \frac{5}{9}}{1 + 2 + 3 + 4} +$ $\frac{3 \times \frac{3}{6} + 4 \times \frac{1}{3}}{1 + 2 + 3 + 4} =$ 0.4528	$1 - 3 \times (\frac{1}{3})^2 = \frac{2}{3}$	$-\frac{1}{4} \times \log_2 \frac{1}{3} = 0.396$
	$\frac{5}{6}$	$\frac{\frac{1}{4} + 1 + \frac{1}{4} + 1 + \frac{3}{4}}{6} =$ $\frac{13}{24}$	$\frac{1}{4}$	$\frac{1 \times \frac{5}{6} + 2 \times \frac{3}{12}}{1 + 2 + 3 + 4} +$ $\frac{3 \times \frac{2}{8} + 4 \times \frac{1}{4}}{1 + 2 + 3 + 4} =$ 0.2562	$1 - (\frac{1}{2})^2 - 2 \times (\frac{1}{4})^2 = \frac{5}{8}$	$-\frac{1}{4} \times$ $(\frac{1}{2} \log_2 \frac{1}{2} +$ $\frac{1}{2} \log_2 \frac{1}{4}) =$ 0.375
	$\frac{10}{10} = 1$	$\frac{\frac{1}{4} + \frac{1}{4} + \frac{2}{4} + \frac{3}{4} + \frac{2}{4}}{10} +$ $\frac{\frac{1}{4} + \frac{2}{4} + \frac{1}{4} + \frac{2}{4} + \frac{1}{4}}{10} =$ $\frac{16}{40} = \frac{2}{5}$	$\frac{4}{5}$	$\frac{1 \times \frac{14}{20} + 2 \times \frac{7}{15}}{1 + 2 + 3 + 4} +$ $\frac{3 \times \frac{3}{10} + 4 \times \frac{1}{5}}{1 + 2 + 3 + 4} = \frac{1}{3}$	$1 - 5 \times (\frac{1}{5})^2 = \frac{4}{5}$	$-\frac{1}{4} \times \log_2 \frac{1}{5} = 0.58$

lation patterns (Figure 5.3A,B). The intra-molecule heterogeneity scores PDR and MHL were less accurate in defining THRs. Since Epipolymorphism and Entropy only consider sequencing reads with at least four CpGs, they were limited in their ability to quantify WSH in CpG-sparse regions. Thus, they could be quantified in only 70 out of the 1,000 regions compared to FDRP/qFDRP which were quantifiable in 912 regions (Figure 5.4). Potential reasons for a non-quantifiable region include poor read coverage in repetitive elements, where sequencing reads cannot be reliably mapped to the reference genome and particularly CpG-sparse regions. Additionally, the WSH scores and the DNA methylation value have been computed using different scripts and software packages, which causes small differences in the number of regions assayed. We found positive Spearman correlation coefficients of the average WSH scores per region and the number of simulated cell types for FDRP, qFDRP and Epipolymorphism (Figure 5.5) indicating that WSH scores reliably quantify the degree of heterogeneity. We did not find a similar correlation of the average DNA methylation level with the simulation parameters, suggesting that the WSH scores are better suited to capture the degree of heterogeneity compared to the DNA methylation level alone.

Currently, epigenomic studies particularly focus on cell-type-specific DNA methylation patterns, which require methods for separating different cell types (e.g, FACS sorting). While these methodologies are well-established, they are also moderately error-prone, which leads to the contamination of samples by non-target cell types. To study the effect of cellular contamination on WSH scores, we simulated data representing contamination by non-target cell types. FDRP, qFDRP, Epipolymorphism, and Entropy were elevated within the THR when we introduced between 0 and 50% cell-type contamination *in silico*. In contrast, MHL was consistently high throughout the regions for some of the simulated datasets, resulting in low AUC values. FDRP and PDR showed a bias for higher values in CpG-dense regions (Figure 5.3). Similarly, the average DNA methylation level reliably detected cellular contamination. In accordance to the cell-type heterogeneity simulation, Epipolymorphism and Entropy were limited in their ability to quantify heterogeneity across all regions (Figure 5.4). In general, MHL and PDR showed

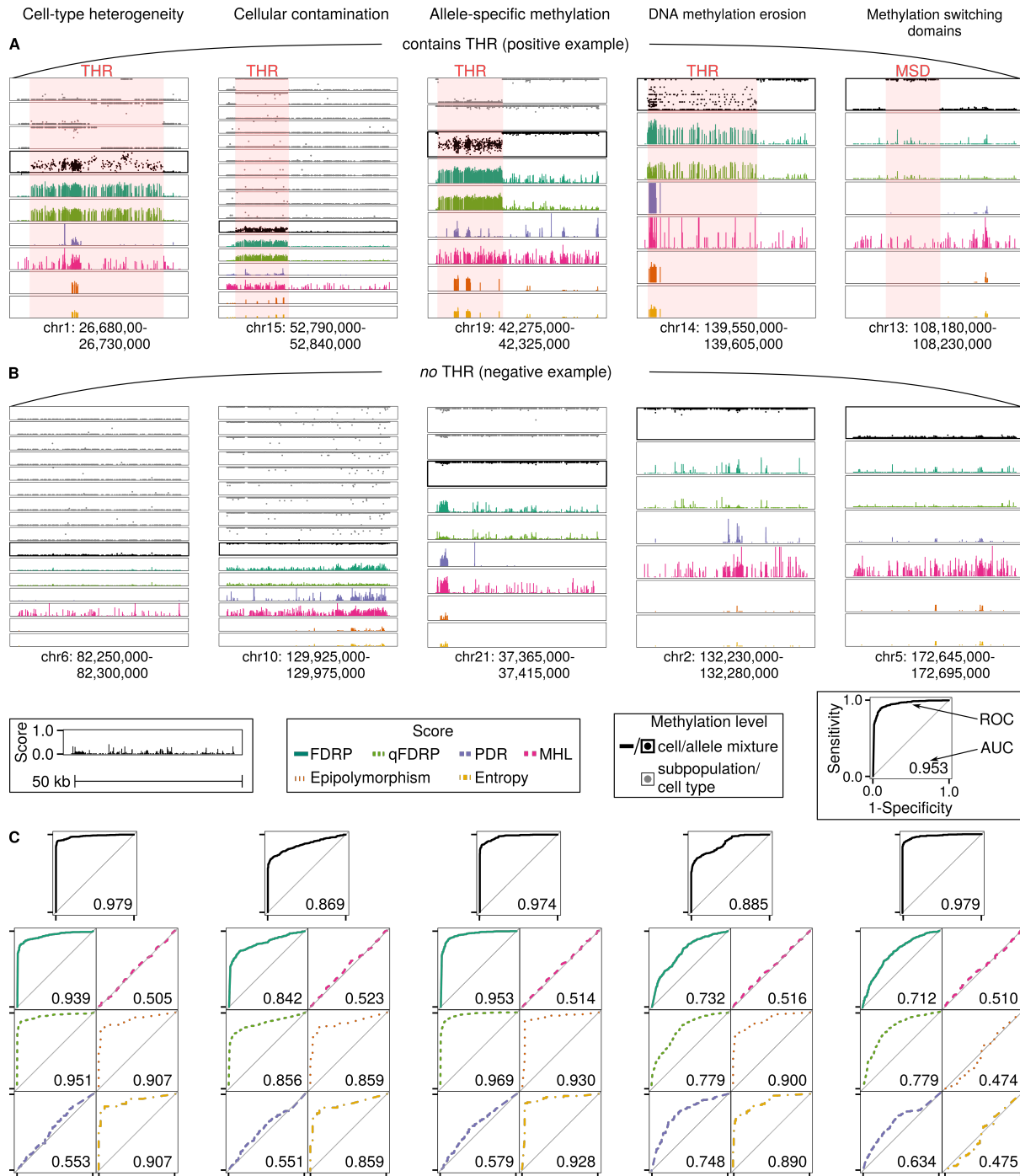


Figure 5.3: WSH scores in five simulation scenarios (Continued on next page).

low accuracy in differentiating between THR and background. As expected, mean WSH scores negatively correlated with the simulated sample purity level using FDRP, qFDRP, Epipolymorphism, and Entropy (Figure 5.5), but not for DNA methylation, PDR, and MHL. Due to the negative correlation with the sample purity level, we speculated that WSH scores could be used as estimates of sample or tumor purity, which we elaborate further below. All reported correlations between the WSH scores and the simulation parameters were significant with respect

Figure 5.3: (Previous page) WSH scores in five simulation scenarios: cell-type heterogeneity, cellular contamination, ASM, DNA methylation erosion, and methylation switching domains. A positive (**A**) and negative (**B**) example selected from the 1,000 simulated regions (size 50 kb) are shown as snapshots for the scores and DNA methylation levels of single cell types (gray points) and cellular mixture (bold outlines) for each scenario. **C:** ROC curves represent the results of t-tests that compare the score/DNA methylation inside THRs with the background across all of the 1,000 simulated regions per scenario.

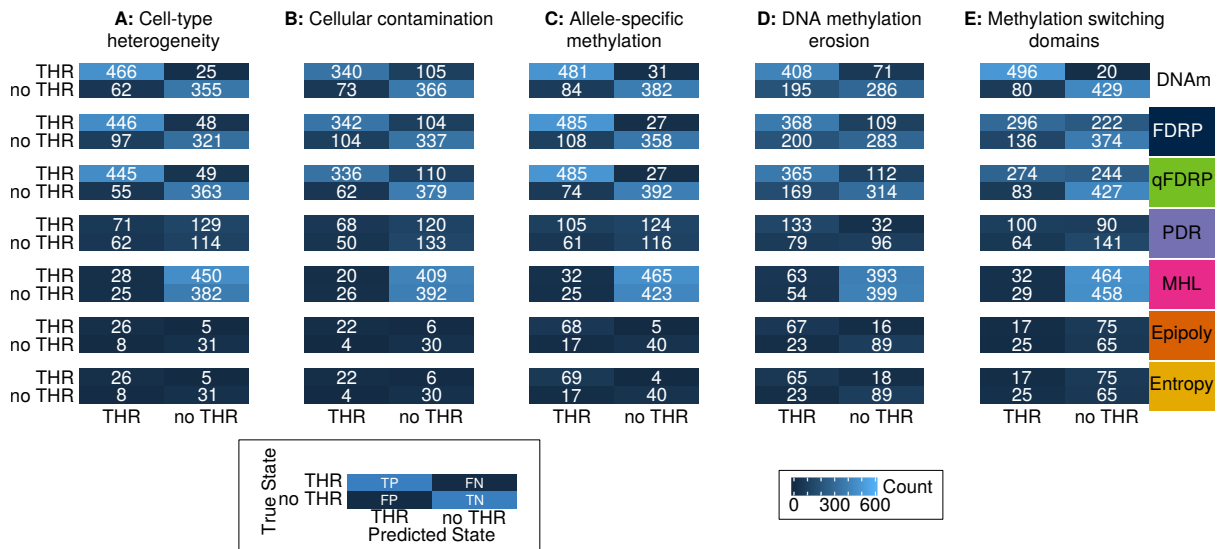


Figure 5.4: Confusion matrices for the five simulation scenarios (**A:** cell-type heterogeneity, **B:** cellular contamination, **C:** allele-specific methylation, **D:** DNA methylation erosion, **E:** methylation switching domains) for all WSH scores and the average DNA methylation level (DNAm). For the confusion matrices, 0.01 was employed as the p-value cutoff to determine if the score detects the THR.

to a two-sided correlation test.

Mammalian genomes are typically diploid, which introduces additional complexity in DNA methylation patterns beyond cell-type heterogeneity. Additionally, ASM has been associated with allele-specific gene expression [285]. In order to evaluate the scores' capabilities to detect regions in the genome showing ASM, we simulated two artificial cell types (here representing alleles) at a fixed 1:1 ratio. Notably, this scenario is a special case of the 'cell-type heterogeneity' and 'cellular contamination' setting, in which only two cell types are mixed at equal proportions. This scenario also models strand-specific methylation, since in most sequencing libraries allele- and strand-specific methylation cannot be differentiated. Except for MHL and PDR, the WSH scores as well as the DNA methylation level accurately identified the THR in the majority of the 1,000 simulated regions (Figure 5.4).

DNA Methylation Erosion During the replication of DNA, a completely unmethylated daughter strand is synthesized based on the template strand. The DNA methylation maintenance enzyme DNMT1 is responsible for copying the DNA methylation state from the template to the newly synthesized strand. During multiple rounds of cell division, cells stochastically lose DNA methylation, since DNMT1 occasionally fails to copy the DNA methylation information

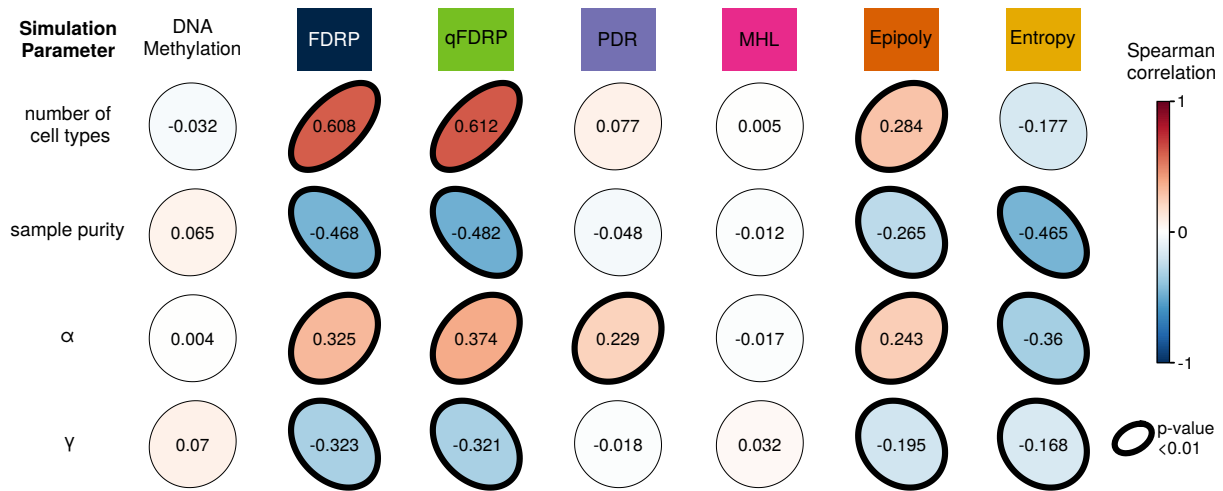


Figure 5.5: Spearman correlation coefficients and associated p-values (two-sided correlation test) between the WSH scores and the simulation parameters: number of cell types, sample purity score, as well as α (stochasticity) and γ (replication) from the DNA methylation erosion scenario. Ellipses are directed towards the upper right for positive and to the lower right for negative correlations, respectively. The color represents the magnitude of correlation. WSH scores were averaged over all CpG positions in each of the regions and this score was correlated to the simulation parameter used for this particular region. Bold outlines of the ellipses indicate significant correlations according to a two-sided correlation test.

to the daughter strand. We refer to this process as DNA methylation erosion and simulated this process by introducing stochasticity in DNA methylation patterns in particular subregions. The inter-molecule heterogeneity scores, although being designed to capture cell-type heterogeneity, also capture DNA methylation erosion in around two thirds of the simulated regions (Figure 5.4). PDR, which is designed for detecting DNA methylation erosion, performed more accurately than in the cell-mixture scenarios and should be highest when the simulation parameter α is close to 50 and for low γ values. Consequently, the stochasticity parameter α quantifying the degree of DNA methylation erosion correlated positively with PDR, but also with FDRP, qFDRP, and Epipolymorphism (Figure 5.5). However, we detected negative correlations between γ , the replication parameter specifying how often a particular pattern is found in the reads, and FDRP, qFDRP, and Epipolymorphism (Figure 5.5), but not for PDR.

Methylation Switching Domains In the last simulation scenario, we aimed at showing that WSH scores are particularly useful for quantifying complex DNA methylation patterns instead of identifying regions or domains with distinct DNA methylation levels in comparison to the background DNA methylation state. Given that the average DNA methylation level could be used to accurately detect THRs in the above scenarios, we tested whether WSH scores specifically capture heterogeneity rather than switches in the DNA methylation level. This means that WSH scores should describe the read configurations in scenarios 2-4 (Table 5.3), which the DNA methylation level fails to capture. Therefore, we assessed each score's performance in detecting methylation switching domains (MSDs), i.e., regions that change the methylation state from fully methylated to unmethylated or vice versa. Such MSDs are typically located at boundaries of active regulatory elements, where a fully methylated background methylation state is

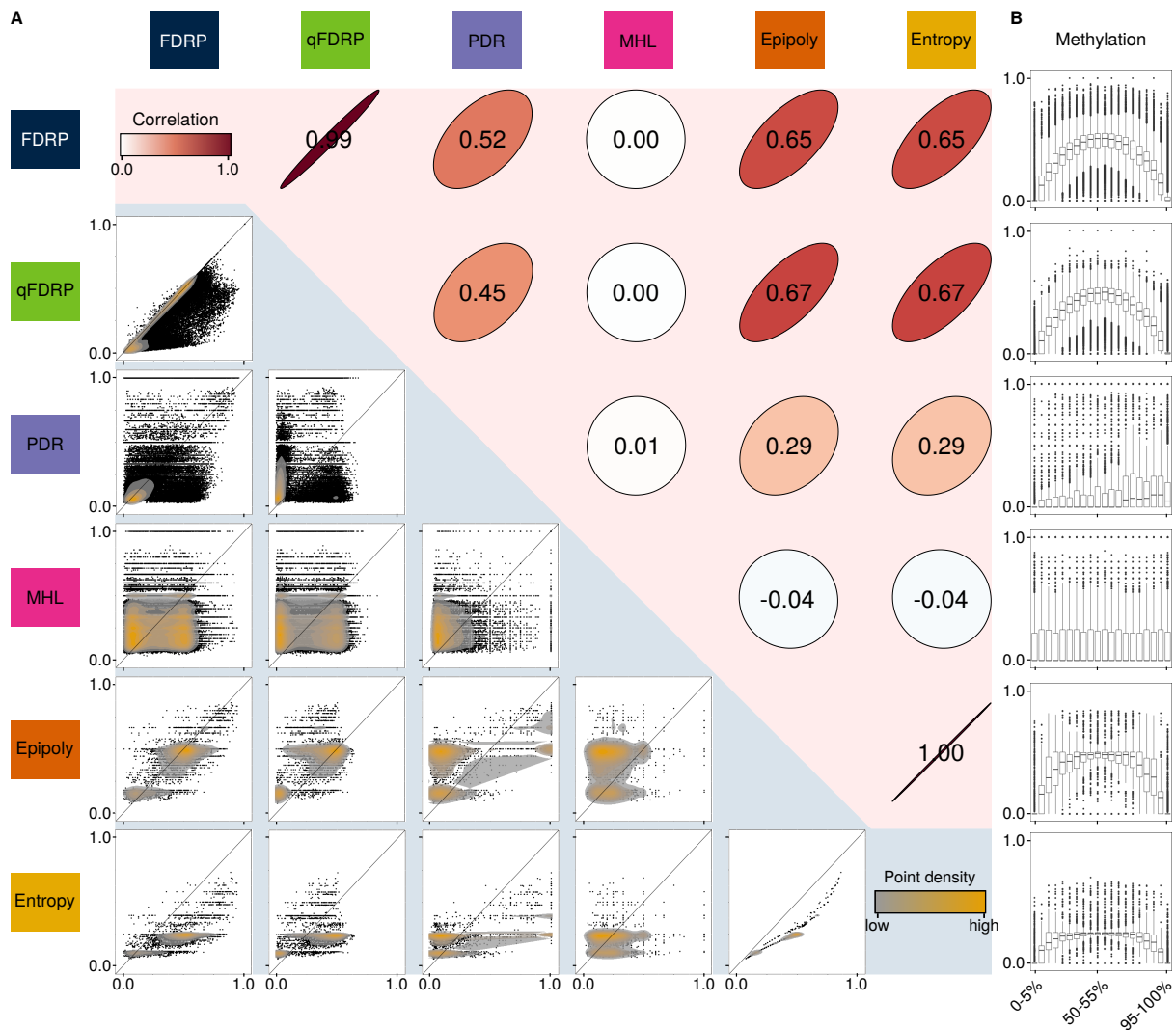


Figure 5.6: Pairwise comparison of WSH scores on simulated data (Continued on next page).

locally reversed to make the regulatory element accessible for TFs. We simulated a single cell type and expected low WSH scores as MSDs do not represent DNA methylation heterogeneity. Consistent with this expectation, we observed a substantially inflated false-negative rate (Figure 5.3, Figure 5.4) for all WSH scores and not for the DNA methylation level. This illustrates that WSH scores indeed contribute additional information to the DNA methylation level.

Shared Information Between Inter- and Intra-Molecule WSH Scores

In order to quantify similarities between the different WSH scores, we merged all regions from the scenarios above and conducted all pairwise comparisons of the WSH scores (Figure 5.6A). FDRP, qFDRP, Epipolymorphism, and Entropy correlated to some extent with the intra-molecule WSH score PDR. This indicates that regions exhibiting locally disordered DNA methylation also show large heterogeneity in sequencing reads. We observed correlation coefficients of 0.65 and 0.67 between FDRP/qFDRP and Epipolymorphism/Entropy, which indicates that the two groups of inter-molecule heterogeneity scores largely describe similar aspects of the DNA

Figure 5.6: (Previous page) Pairwise comparison of WSH scores on simulated data. **A:** Comparison between the WSH scores using all datasets from the four heterogeneity simulation scenarios. Blue triangle: Scatterplots comparing WSH scores. Each point is a CpG site or four CpG window for which both scores quantified WSH. High-point-density areas are visualized using yellow and grey was used for low-density regions (density estimated through kernel density estimation). Points with values zero for both of the scores were removed from the scatterplots for better visualization (retained in the original publication [268] and for computing the correlation coefficients). Red triangle: Spearman correlations between the scores. **B:** Dependency on DNA methylation. DNA methylation levels were binned in steps of 5% methylation, and the CpG-wise WSH score (y-axis) was compared to the DNA methylation level (x-axis).

methylation landscape. However, there are also distinct regions showing differences across the scores. Additionally, Epipolymorphism/Entropy captured substantially fewer regions than qFDRP/FDRP (Figure 5.4), which capture new regions not yet considered by existing metrics. MHL was unrelated to the other scores. Except for MHL and PDR, the WSH scores were generally higher in intermediately methylated regions than in completely methylated or unmethylated regions (Figure 5.6B).

Robustness Regarding Technical Biases

Systematic biases in data due to the measurement technology constitute an important challenge for data analysis [286] and the WSH scores discussed here should be independent of several technical parameters. Thus, we systematically simulated differences in technical setup and genomic constitution including read coverage, read length, sequencing errors, and CpG density.

To investigate the effect of sequencing depth on WSH scores, we generated bisulfite sequencing data at different read depths (between 5,000 and 50,000) for each of 1,000 randomly selected regions of size 50 kb individually. Epipolymorphism and Entropy required rather high coverage (Figure 5.7A), while the other scores quantified heterogeneity even in low coverage datasets. FDRP, PDR, Epipolymorphism, and Entropy increased with higher read coverage and MHL/qFDRP were independent of sequencing depth. Additionally, we investigated the relationship between CpG-wise read coverage and the WSH scores in an individual dataset (i.e., a single region from our simulation scenarios) by computing Spearman's rank correlation between the CpG-wise number of reads and the WSH score at this CpG. All WSH scores were independent of CpG-wise read coverage. Thus, all scores can be used to compare heterogeneity in different regions with potentially different coverages, but not to compare datasets with different genome-wide coverage.

Sequencing read lengths can differ between datasets generated in different contexts or with different protocols. Thus, we systematically investigated the dependency between WSH scores and read length. We chose read lengths between 40 and 150 bp, which cover the most frequently applied (Illumina) sequencing methods¹⁰. PDR and FDRP increased with longer reads (Figure 5.7B), while qFDRP, MHL, Epipolymorphism, and Entropy were independent of read length. Notably, Epipolymorphism and Entropy struggle with read lengths below 80 bp. Since reads shorter than 50 bp are rarely used, qFDRP and MHL can be applied to datasets of any read

¹⁰<http://www.biotech.cornell.edu/brc/genomics-facility/services/next-generation-sequencing>

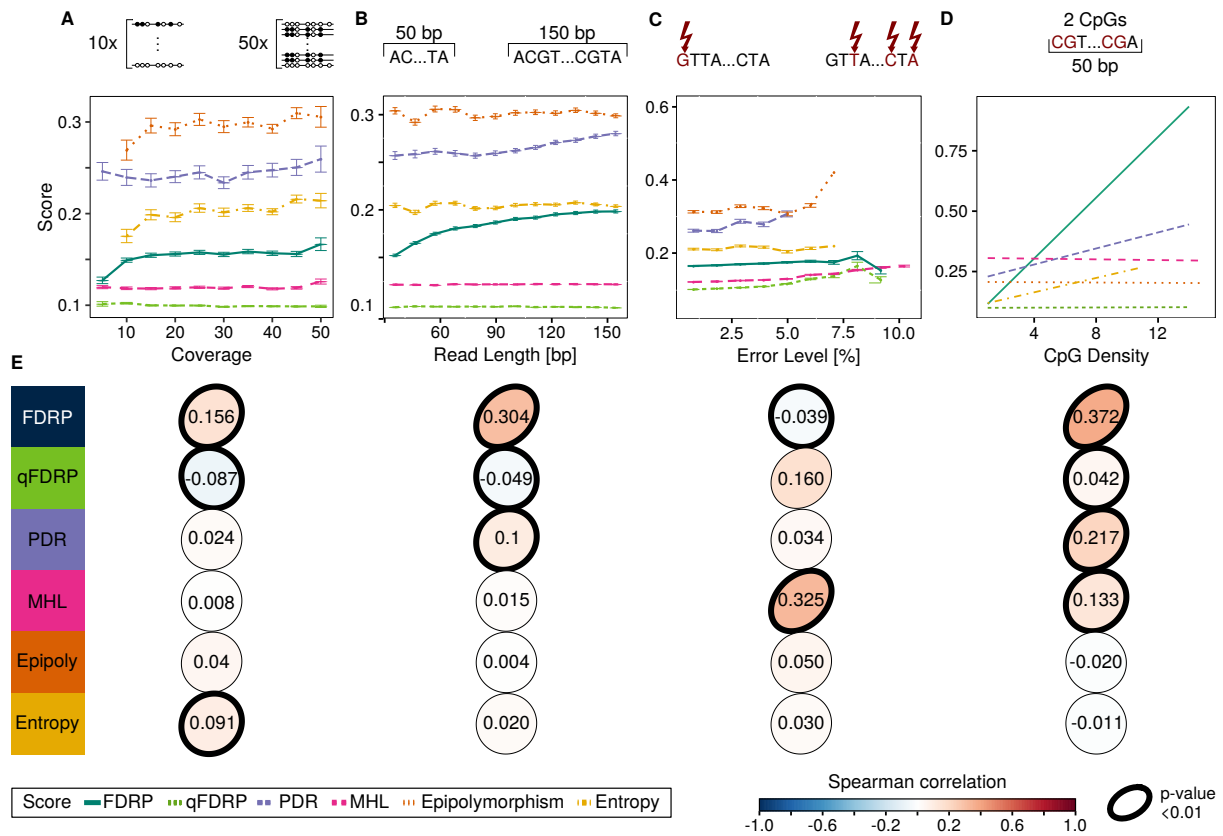


Figure 5.7: Influence of technical biases on WSH scores. **A:** Average WSH scores vs. average coverage. The error bars indicate standard errors in different simulated regions of the same average coverage. **B:** Average WSH scores and standard errors vs. length of the simulated reads. Average WSH scores and standard error vs. **C** simulated sequencing error level in percent and **D** the number of CpGs in 50 bp windows (CpG Density). CpG density is a quantitative variable in contrast to coverage, read length, and sequencing error level, which is why linear regression lines are shown instead of lineplots. **E:** Spearman's rank correlation between the WSH scores and the simulation parameters depicted in A-D. Significant (at level 0.01) p-values according to two-sided correlation tests are depicted with a bold outline of the ellipse.

length and potentially also for comparing datasets with different read lengths. Using any of the other scores, the influence of read length should be thoroughly investigated when integrating multiple datasets.

Similarly, we investigated the susceptibility of the scores to sequencing errors. FDRP and qFDRP quantified heterogeneity up to an error level of 9% and are not applicable to datasets that are more error-prone (Figure 5.7C). Epipolymorphism and Entropy support an error rate of at most 7%, while PDR only quantifies WSH up to 5% error level. With increasing error level, effective sequencing coverage decreases, since fewer reads can be reliably aligned to the reference genome. Thus, more regions/CpGs will be discarded from the analysis using a coverage cutoff, which especially affects the scores requiring four CpGs per sequencing read. MHL, qFDRP, Epipolymorphism, and Entropy were rather stable up to 5% error level and beyond that sequencing errors were incorrectly considered as heterogeneity. In summary, all scores are sensitive to sequencing errors, but can cope with error rates up to 5%, which is above error percentages detected in Illumina sequencing (0.5-2%, [287]).

Table 5.4: WSH statistics computed on the healthy blood dataset (blood cohort), the *in-silico* mixed WGBS samples (DEEP hybrid samples), and the cancer example (Ewing) with number of rows (sites/regions) and percentage of missing values (NAs) for all considered WSH scores.

WSH Score	blood cohort (RRBS)		DEEP (WGBS)		Ewing sarcoma (RRBS)	
	# sites	% NAs	# sites	% NAs	# sites	% NAs
FDRP	1,176,471	2.11%	24,198,968	38.81%	1,227,943	5.75%
qFDRP	1,176,471	1.7%	24,198,968	38.81%	1,227,943	5.75%
PDR	1,176,471	62.31%	24,198,968	76.66%	1,227,943	65.82%
MHL	388,848	19.88%	4,590,846	0%	333,542	27.28%
Epipolymorphism	549,129	73.1%	740,216	0%	697,022	83.1%
Entropy	549,129	73.41%	740,216	0%	697,022	83.1%

To model the connection between WSH scores and genomic base composition, we calculated local CpG density for each of the simulated regions as the number of CpG dinucleotides in a 50 bp window. Then, we computed Spearman's rank correlation between the average WSH score in the 50 bp window and the number of CpGs. FDRP and PDR, but also MHL, correlated with CpG density (Figure 5.7D). This is likely caused by the rigid classification of each read/read pair as discordant even if only a single methylation state differs from all the others. Thus, the probability of detecting a mismatch of DNA methylation states increases when investigating more CpG sites per read both for PDR and FDRP, which is the case for CpG-dense regions. This was the original motivation for developing qFDRP, which accounts for this dependency and is independent of local CpG density as are Epipolymorphism and Entropy.

Confirmation of Simulation Results on Bisulfite Sequencing Data

We validated the findings obtained in the simulation experiments on a human longevity cohort of healthy individuals, comprising 239 whole blood samples assayed using RRBS. Notably, computing MHL using the script accompanying the original publication [277] required the longest runtime among the scores (average wallclock runtime 35 hours per sample vs. 1:30 hours for qFDRP/FDRP, Supplementary Figure A.7). Spearman correlations between the scores, especially between intra- and inter-molecule scores, were higher than those on simulated data. This further emphasized that locally disordered methylation and local variation in DNA methylation patterns coincide. Epipolymorphism and Entropy described largely the same heterogeneity as qFDRP and FDRP, which leads to high correlations in the regions that are captured by the two types of scores. However, qFDRP and FDRP captured more than twice the number of regions in comparison to Epipolymorphism/Entropy (Table 5.4). Unexpectedly, Epipolymorphism was negatively correlated with average coverage per sample and MHL was strictly bimodally distributed similar to the overall DNA methylation level. We found that heterogeneity was unevenly distributed across the genome and was preferentially located in distal, rather than proximal regulatory elements defined by the Ensembl Regulatory Build [8] (Figure 5.8).

To show that WSH scores capture WSH independent of the sequencing technology used, we constructed a homogeneous and a heterogeneous sample *in-silico* using WGBS data from liver samples assayed within the DEEP project. All scores, except for MHL, exhibited elevated heterogeneity in the heterogeneous sample. The analysis also validated that WSH is preferentially

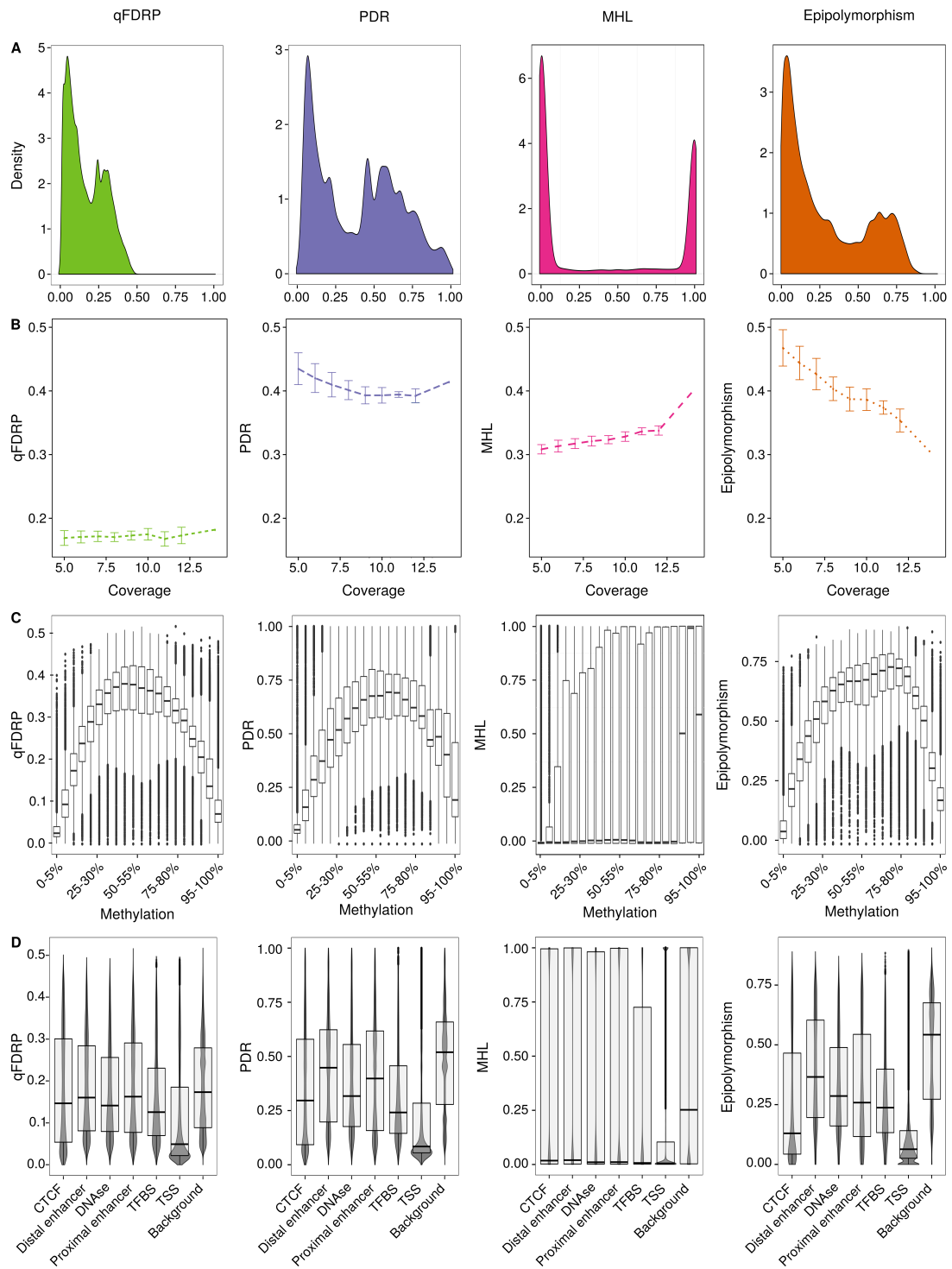


Figure 5.8: WSH scores in the blood cohort dataset (Continued on next page).

located in distal rather than proximal regulatory elements (Supplementary Figure A.8).

Figure 5.8: (Previous page) WSH scores in the blood cohort dataset. **A:** Genome wide-distribution of WSH scores. **B:** Per-sample average WSH score vs. per-sample average coverage. **C:** WSH scores vs. DNA methylation stratified into 20 classes of size 5% each. **D:** Distribution of WSH scores across putative regulatory elements defined by the Ensembl Regulatory Build are shown for qFDRP, PDR, MHL, and Epipolymorphism.

Differentially Heterogeneous Regions in Cancer

WSH scores could be especially relevant for quantifying intra-tumor heterogeneity in solid tumors. Thus, we quantified DNA methylation heterogeneity in a Ewing sarcoma dataset (Figure 5.9), comprising different types of samples including Ewing tissue, Ewing cell lines (CL), and mesenchymal stem cells (MSC) as the potential cell-of-origin population for Ewing tumors [288]. MSCs were further stratified according to the health state of the donor into eMSCs, which originate from Ewing sarcoma patients [154], or normal MSCs, respectively. We elaborate on an example analysis using qFDRP, but analogous analyses using the other WSH scores can be found in the original publication [268]. In summary, results obtained using Epipolymorphism and Entropy were similar to qFDRP, while qFDRP quantified substantially more regions (Table 5.4). When we compared Ewing sarcoma samples to the samples from the blood cohort, qFDRP indicated higher overall heterogeneity in the set of cancer samples (Figure 5.9A). In particular, we detected highest heterogeneity in MSCs and slightly higher values in eMSCs (mean: 0.229) compared to normal MSCs (mean: 0.214). qFDRP was lower in the Ewing CLs (mean: 0.185) and Ewing tissue samples (mean: 0.184). In accordance to the findings on the blood cohort, lowest WSH was detected in TSS, which is also in line with the lowest average DNA methylation level.

Differences in tumor purity levels constitute another important challenge for computational analyses of cancer samples. Thus, estimates of tumor purity are critical for downstream analysis in cancer research. Tumor purity levels are typically obtained from histopathological investigations, but can also be estimated from genetic or epigenetic data [149]. We tested whether WSH scores can be used to reliably predict tumor purity levels. To do so, we trained a Lasso model to select 26 sites significantly associated with annotated tumor purity levels for a subset of the samples. Using qFDRP, we could demonstrate good prediction performance with an overall cross-validated mean absolute error of 0.027 at a correlation of 0.966. Then, we employed this model for the prediction of tumor purity levels for the 48 samples without prior annotation. We found that the samples clustered together with those samples showing a similar level of annotated tumor purity (Figure 5.9). In general, we observed that the sample clustering on the CpG sites associated with the tumor purity level was mainly associated with the tumor purity levels as expected. Additionally, qFDRP values of the selected sites were significantly higher in the low purity cluster than in the high purity cluster (mean 0.133 vs. 0.291, two-sided t-test p-value lower than 2.2×10^{-16}). Lastly, we repeated the analysis using random sample permutations and found that results were poor (cross-validation-correlation: 0.02, cross-validation mean absolute difference: 0.13), demonstrating that our model captured tumor purity.

To determine CpGs associated with the development of Ewing sarcoma, we performed a differential WSH analysis between MSCs and Ewing tissue samples using *limma* [246]. In accordance with the genome-wide mean, MSCs (both normal MSCs and eMSCs) exhibited higher

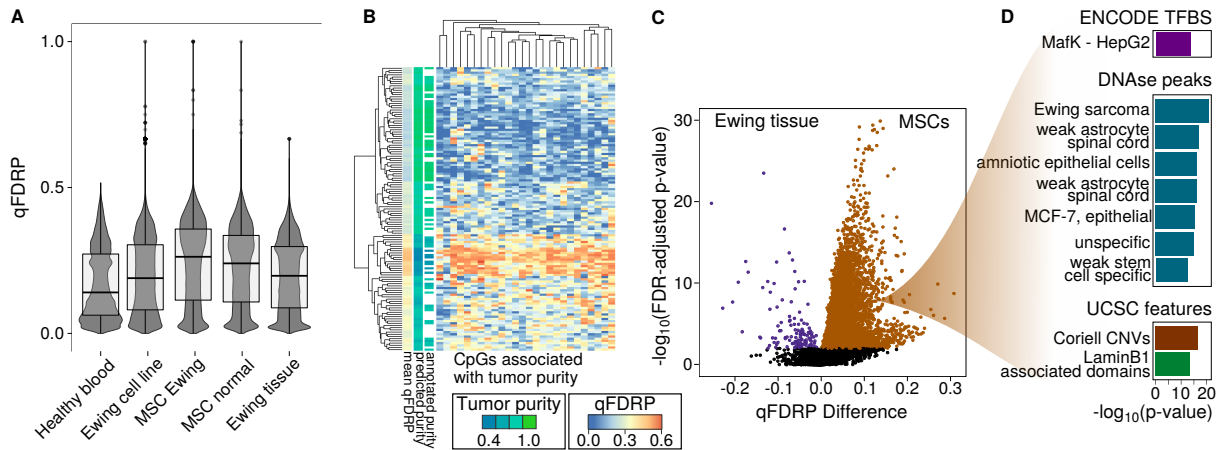


Figure 5.9: WSH in Ewing sarcoma samples. **A:** Combined box-violin plot of qFDRP values for different sample groups. **B:** Heatmap (blue low, red high value) of CpG-wise qFDRP scores (Ewing tissue samples without FFPE samples) for the 26 sites associated with tumor purity. Both samples (rows) and CpGs (columns) were hierarchically clustered (Euclidean distance, complete linkage). The annotated color for the samples indicates predicted and annotated tumor purity levels, as well as the average qFDRP value across the selected sites. **C:** Volcano plot of qFDRP values aggregated along gene bodies in Ewing tissue samples versus MSCs. Positive values on the x-axis indicate higher WSH in MSCs. Each point is a gene, which is color-coded if it has FDR-adjusted p-value not more than 0.01. **D:** LOLA enrichment analysis of MSC-hyper-heterogeneous genes. Histograms represent the negative common logarithm of the enrichment p-value. MCF-7, breast cancer cell line; CNV, copy number variation

heterogeneity than tissue samples also after aggregation along annotated gene bodies. We found that genes that had higher average qFDRP in MSCs compared to Ewing tissue (hyper-heterogeneous in MSCs; Figure 5.9C) were preferentially located in DNaseI-hypersensitive regions in various cell types. Notably, those genes were also enriched for DNaseI-hypersensitive sites associated with Ewing sarcoma in the original publication of the Ewing dataset [154] for qFDRP and the DNA methylation level (Figure 5.9D). Finally, differentially heterogeneous and differentially methylated genes were preferentially located in different TFBS: qFDRP enriched for MafK, and differential DNA methylation levels were detected in binding sites of c-MYC, c-FOS, and GATA3. The differential analysis further emphasized that WSH scores capture a distinct aspect of the DNA methylome.

5.1.4 Discussion

We benchmarked six scores created for quantification of within-sample heterogeneity in both simulation experiments and bisulfite sequencing datasets. Using simulated data, we showed that the power of WSH scores to detect heterogeneity varied depending on their design motivation. For instance, PDR did not capture inter-molecule heterogeneity, since it was created to capture locally discorded methylation, i.e., intra-molecule heterogeneity. However, PDR did not outperform the inter-molecule heterogeneity scores, when we simulated DNA methylation erosion as an instance of intra-molecule heterogeneity. Thus, it remains elusive, which distinct source of DNA methylation heterogeneity is addressed by PDR. MHL mimics DNA methylation haplotype blocks in bisulfite sequencing reads and might thus not be suitable for capturing

WSH as defined in this work, but rather describes a distinct aspect of the methylome. PDR, Epipolymorphism, and Entropy require four CpGs per sequencing read by definition, which potentially masks regions of low CpG content. Further studies are required to show the applicability of Epipolymorphism and Entropy using different numbers of CpGs per epiallele. The score that we propose, qFDRP, is not limited by definition. The inter-molecule heterogeneity scores were relatively highly correlated with the intra-molecule heterogeneity scores, and we speculate that, in order to exhibit large variations in the sequencing reads, the reads themselves need to be locally disordered. Thus, inherently heterogeneous regions show both high variance in sequencing patterns, but also locally disordered methylation.

Since sequencing reads shorter than 50 bp are rarely used, qFDRP, MHL, Epipolymorphism, and Entropy can be applied to most published datasets irrespective of the read length and also for comparing datasets employing different read lengths. Notably, all scores were sensitive to sequencing errors in our simulations, but tolerated error levels up to 5%. This level is above the error percentages reported for Illumina sequencing (0.5-2% [287]). This analysis could potentially be extended to datasets beyond Illumina sequencing, such as short read sequencing by BGI¹¹, or long-read (third-generation) sequencing by Oxford Nanopore or PacBio. For long-read sequencing, important parameters of the WSH scores have to be adapted, such as considering four consecutive CpGs in Epipolymorphism, Entropy, and PDR. Additionally, it is necessary to investigate whether the current coverage of long-read sequencing is sufficient for computing WSH scores or whether the substantially higher error levels make the quantification of WSH infeasible. We note that we did not systematically investigate the influence of experimental batch effects, such as different laboratories, restriction enzymes, or differences in the genomic coverage of WGBS and RRBS. Additionally, PCR duplication artifacts were not part of our simulations and remain to be investigated. These duplicates are generally removed during low-level data processing for WGBS data (see Section 2.5.2), but remain hard to detect for RRBS data.

As a validation of the simulation results and to show potential applications in a clinical setting, we analyzed three experimental bisulfite sequencing datasets. Using tools employed for identifying differential DNA methylation levels, such as *limma*, we were able to quantify differential heterogeneity between groups of samples. We detected higher WSH in the MSCs used here in comparison to tumor samples. Stem cells represent heterogeneous populations of cells [289], which leads to high WSH when analyzing multiple cells from a pool of heterogeneous stem cells. In contrast, tumor cells follow a more clonal behavior. Another explanation for the elevated WSH could be technical issues in sample preparation. Lastly, DNA methylation oscillations were reported in primed ESCs, which originate from increased expression of DNMT3A/B together with high expression of TET enzymes [290]. DNA methylation erosion in clonal populations of tumor cell lines has recently emerged as an important biological observation [291] and PDR, but also qFDRP, could be used to systematically investigate DNA methylation erosion in larger cohorts. We reported that qFDRP can reliably predict tumor purity levels estimated from genetic data, which can be valuable if such data is missing. As expected, higher heterogeneity was detected in those samples that had lower tumor purity estimates. It remains to be further investigated whether qFDRP is able to predict tumor purity levels in datasets other than Ewing sarcoma.

WSH was preferentially located in regions not yet annotated to a functional category accord-

¹¹<https://bgi.com/global/>

Table 5.5: General guidelines for the application of WSH scores in epigenomic studies.

Score	Concept	Strengths	Drawbacks	Application scenario
PDR	Locally disordered methylation	Detects DNAm erosion CpG-wise score Fast computation	Simulated heterogeneity <i>not</i> detected Dependency on read length and CpG density	Addressing locally disordered DNA methylation in large cancer datasets
MHL	Methylation haplotypes	CpG-wise score Robust to technical setup	Simulated heterogeneity <i>not</i> detected Slow computation	Linking genetically detected haplotypes to DNA methylation haplotypes
Epipoly	Variance among the reads	Simulated heterogeneity detected Robust to technical setup	<i>no</i> CpG-wise score Few regions captured	Segmentation into highly and lowly variably methylated regions for large bisulfite sequencing datasets
Entropy	Variance among the reads	Simulated heterogeneity detected Robust to technical setup	<i>no</i> CpG-wise score Few regions captured	Segmentation into highly and lowly variably methylated regions for large bisulfite sequencing datasets
FDRP	Variance among the reads	Simulated heterogeneity detected CpG-wise score	Dependency on coverage, read length, and CpG density Rather slow computation	Linking CpG-wise methylation values to epigenetic heterogeneity in large bisulfite sequencing datasets
qFDRP	Variance among the reads	Simulated heterogeneity detected Robust to technical setup CpG-wise score	Rather slow computation	Linking CpG-wise methylation values to epigenetic heterogeneity in large bisulfite sequencing datasets

ing to the Ensembl regulatory build in all three bisulfite sequencing datasets in line with the results presented by Feinberg et al. [292]. These regions might be missed using established techniques such as the average DNA methylation level. Their functional role and connection to diseases warrant further investigation. We envision that WSH scores can be used to segment the genome into regions with particularly high or low heterogeneity, similar to the definition of PMDs [18, 153]. Since reduced correlation of neighboring CpGs has been reported in cancer [293], PDR is a premier candidate for more detailed investigations. Since we expect cell-type heterogeneity to be the major driver of WSH, the scores proposed here could be used in combination with cell-type deconvolution tools, such as *MeDeCom* (see Chapter 4).

Recommendations and Guidelines

Table 5.5 summarizes strengths and limitations of WSH scores in capturing different characteristics of heterogeneity. PDR quantifies locally disordered regions in large cancer datasets, but is dependent on data quality and genomic features, especially read length and CpG density.

Additionally, since by definition four CpGs per sequencing read are required, application on datasets with short reads (e.g., 50 bp) is not recommended. The restriction to four CpGs is especially critical, since CpG dinucleotides are heavily depleted throughout the genome. What is more, CpG-dense regions (i.e., CGIs), which co-localize with promoters, generally show lower degrees of WSH and thus less dynamic behavior across different samples is expected. In contrast, we found elevated qFDRP in distal regulatory elements, including putative enhancers, which points toward the applicability of qFDRP to detect novel regulatory regions. Notably, one could use PDR without restricting only on reads with at least four CpGs. PDR and FDRP were sensitive to technical setup because of the strict classification of each read (pair) as discordant/concordant. qFDRP is particularly suitable for identifying regions exhibiting high heterogeneity due to cell-type differences and complements CpG-level DNA methylation measurements. It also proved to be robust with respect to technical noise in our simulation setup and was independent of sequencing coverage in the experimental datasets. Epipolymorphism and Entropy are suitable for region-based analysis while they fail to capture heterogeneity in CpG-sparse regions, since they are restricted to regions with at least four CpGs per read. MHL was less specific in quantifying WSH. While it was robust to technical variation in synthetic data, it did not correlate to the DNA methylation level and to the other scores.

WSH scores provide insights into sample composition and cell subpopulations. Thus, they complement the DNA methylation level by revealing differences among individual cells and alleles with unknown functional impact. Nevertheless, to date WSH is rarely considered in epigenomic studies. Here, we provide the first systematic and comprehensive evaluation of WSH scores that capture DNA methylation pattern variations or locally disordered methylation directly from the sequenced reads. In contrast to the approach presented in Chapter 4, WSH is used as a feature rather than considered a confounding factor. Based on simulations and experimental data, we provide guidelines for selecting the WSH scores most appropriate for complementing DNA methylation levels as surrogates of heterogeneity. Our results indicate that WSH scores are suitable for the identification of genomic regions in which DNA methylation heterogeneity drives phenotypic changes in development and disease.

Conclusions and Outlook

This last chapter summarizes the results, discusses implications of the presented findings, assesses the impact of the different aspects of the work, and presents potential future directions and extensions. I will show how the tools and workflows presented throughout the thesis can be used to tackle heterogeneity in DNA methylation data at the three levels, and how this can be leveraged to obtain a better understanding of biological regulation. Additionally, potential limitations of the approaches will be discussed. Based on that, I will argue about future directions of the presented work and of epigenomic research in general.

6.1 Summary and Perspectives

Throughout this work, software tools, analysis workflows, and their application on DNA methylation datasets have been presented. More specifically, this thesis addressed heterogeneity in DNA methylation data at three levels: between phenotypes, between individuals sharing a phenotype, and within a sample. In addition to presenting novel software solutions for analyzing DNA methylation data, we applied the tools to available datasets to discuss the biological implications associated with DNA methylation heterogeneity.

Between-group heterogeneity is the best-studied level of DNA methylation heterogeneity. Differences in DNA methylation states associated with a phenotype such as a disease can be leveraged for identifying novel disease biomarkers, which contribute to the improvement of precision medicine. To further the understanding of disease-associated aberrations in DNA methylation, software tools are needed that are easy to use even for non-bioinformaticians. The recent update of the *RnBeads* software package presented in this work provides novel methods for quantifying differential variability, novel covariate inference methods (sex and age), and methods for genome-wide segmentation. We envision that *RnBeads* can be used by many scientists ranging from clinicians to biologists and experienced bioinformaticians. We will continually maintain and extend *RnBeads* to keep the software up-to-date with the developments in the epigenomic research community. New potential features for *RnBeads* include: (i) *de-novo* DMR calling using methods such as *BSmooth* [172] or *DSS* [294], (ii) extension of DNA methylation-based segmentation according to the *MethylSeekR* approach [153], (iii) extending the software package to use read-level information for computing WSH scores, and (iv) an adaptation of the pipeline for supporting single-cell bisulfite sequencing data. *RnBeads* is a software tool that is widely-used in the scientific community, which is indicated by the number of support questions that we answer using the *RnBeads* mailing list¹ and through forum questions, as well as by between 200 and 400 downloads from Bioconductor² each month and

¹team@rnbeads.org

²<http://bioconductor.org/packages/stats/bioc/RnBeads/>

by 16 citations according to Web of Science³ in the year 2020 (status: December 2020). *RnBeads* is available through collaborative software efforts such as *bioconda*⁴ and is used within the de.NBI⁵ and SYSCID⁶ projects. We will continually present *RnBeads* to the scientific community through an extensive documentation and associated website⁷, courses, workshops, and tutorials.

A main driver of DNA methylation differences between phenotypes is genotype variation across the individuals investigated. With the *MAGAR* R-package – the second software application developed in this thesis – we proposed the first software tool that handles raw DNA methylation and genotyping microarray data for detecting associations between genotype and DNA methylation state (methQTLs). *MAGAR* was used in combination with colocalization analysis to investigate tissue-specificity of methQTLs and we found that both tissue-specific and common methQTLs can be detected. Importantly, tissue-specific methQTLs were preferentially located in enhancer elements, which is in line with the important regulatory role of enhancer elements for imprinting cellular identity. We expect that more integrative analysis between genotyping and DNA methylation data will be conducted in the future and expect *MAGAR* to be of significant use for these analysis. An important application of methQTLs is to further the understanding of epigenetic and genetic dysregulation associated with diseases. Associations between a genotype alteration (GWAS) or DNA methylation alteration (EWAS) and a disease are preferentially located in non-coding regions of the genome [218] and the triangular relationship between genome, epigenome, and the disease can be illuminated through combined methQTL, GWAS, and EWAS analysis. Throughout this work, we employed colocalization analysis through SMR analysis to define tissue-specificity, but colocalization methods [295] are readily applicable for unraveling more complex relationships. Similarly, methQTL and eQTL results on healthy individuals can be combined through colocalization analysis to illuminate the relationship between genotype, DNA methylation, and gene expression states [59, 60] (Figure 6.1).

DNA methylation heterogeneity between phenotypes (e.g., EWAS) is closely intertwined with between-sample heterogeneity, since the latter is an important confounding factor for the former through cell-type heterogeneity. Accounting for potentially altered cell-type compositions of samples is critical for determining reliable associations between DNA methylation and a phenotype, otherwise identified DMCs or DMRs may be associated with an altered cell composition rather than with the phenotype of interest. Thus, we recommend to use cell-type-adjustment methods for epigenomic studies. For blood-based studies, reference-based methods such as the Houseman approach [224], *EpiDISH* [225], or *MethylCIBERSORT* [227] return reliable estimates of cellular proportions. However, for insufficiently characterized, complex tissues, such as tumor samples, reference-free methods, including *MeDeCom*, are required. To streamline deconvolution analysis and to provide an easy-to-use pipeline reaching from raw DNA methylation data to the guided biological interpretation of deconvolution results, we presented a three-stage protocol. The protocol comprises (i) data processing, (ii) deconvolution, and (iii) guided biological interpretation of deconvolution results. We applied the protocol to lung adenocarcinoma and melanoma data and identified in both cohorts associations of the

³<https://webofknowledge.com>

⁴<https://bioconda.github.io/recipes/bioconductor-rnbeads/README.html>

⁵<https://www.denbi.de/>

⁶<http://syscid.eu/>

⁷<https://rnbeads.org>

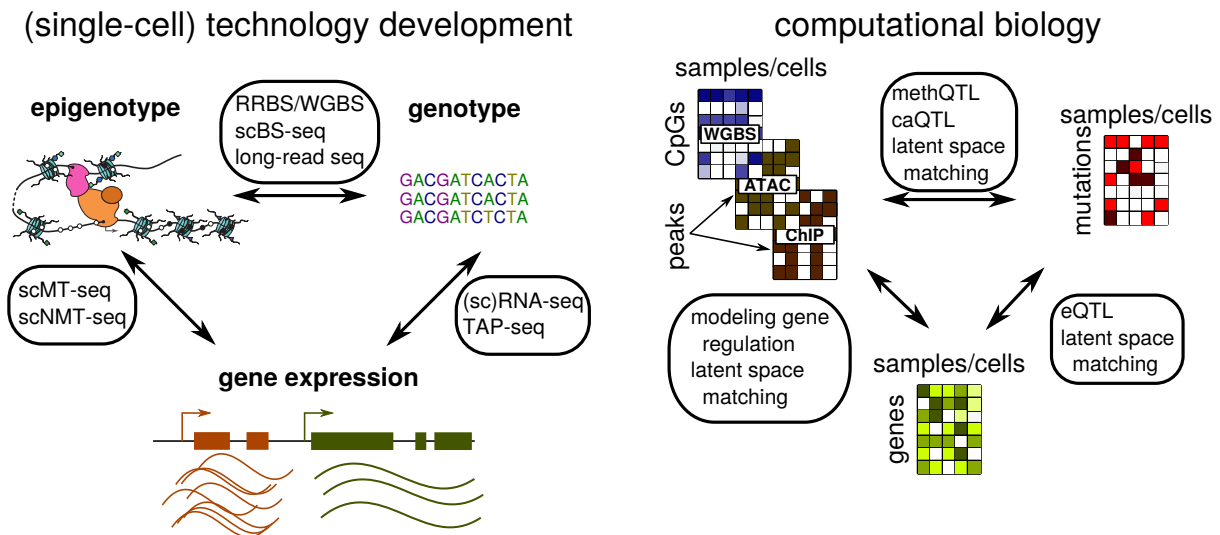


Figure 6.1: Integrating genomic, epigenomic, and transcriptomic data. Data integration of different types of molecular data occurs both on the level of technological developments (left panel), where multiple data layers are measured in the same sample/cell and on the level of computational biology, where the different layers of information are combined (right panel). This applies both to single-cell technologies, where multiple data layers are measured on the same cell and to bulk experiments, where single-molecule information is encoded in the sequencing reads and can be uncovered using computational approaches. The arrows represent the joint characterization of the molecular layers, which can be achieved through the methods listed in the boxes. scBS-seq, single-cell bisulfite sequencing [100]; scMT-seq, single-cell methylation and transcriptome profiling [101]; scNMT, single-cell nucleosome, methylation, and transcription sequencing [296], TAP-seq; targeted pertub-sequencing [297]; modeling gene regulation, predicting gene expression from epigenomic data, TEPIC [10]; latent space matching; constructing a common low-dimensional embedding of different data modalities [298, 299]

detected components with immune cell infiltration into the tumor. Furthermore, the identified LMCs comprised information about patient survival. Similar applications of the protocol to other cancer types can further the understanding of tumor heterogeneity and its implications on patient prognosis and ultimately lead to improved therapy selection. The second stage of the deconvolution protocol employed *MeDeCom*, but the pipeline is readily adaptable to other reference-free and reference-based deconvolution tools. In the last stage of the protocol, the graphical user interface *FactorViz* was used to interpret deconvolution results. Result interpretation still requires user interaction, but we envision that a fully automated interpretation of deconvolution results using statistical learning algorithms can be developed. Additionally, new interpretation features can be easily added to the existing framework.

The majority of epigenomic studies investigates DNA methylation heterogeneity between phenotypes for biomarker discovery, while an increasing number of studies also adjusts for within-group heterogeneity using deconvolution tools. In contrast, the third level of heterogeneity addressed here – within-sample heterogeneity (WSH) – is rarely studied. A potential reason for the lack of epigenomic studies investigating WSH is that the widely-used Illumina microarray does not allow for a genome-wide assessment of WSH, since the generated beta-value is a merged signal across different cellular states. Bisulfite sequencing data affords comprehensively studying WSH, since each sequencing read reflects a distinct cellular state. A

genome-wide assessment of WSH is crucial for gaining insights about the distribution and function of WSH in human methylomes. Within this work, we systematically compared different genome-wide WSH scores and proposed a novel score – qFDRP – with single-CpG resolution. In comparison to existing metrics, qFDRP was less affected by technical parameters and captured different sources of WSH in simulated and experimental datasets. We investigated cell-type heterogeneity, cellular contamination, allele-specific methylation, and DNA methylation erosion as potential sources of WSH, but additional scenarios could be added to the simulations. Using published RRBS and WGBS profiles, we found that WSH is preferentially located in distal regulatory elements, such as DNaseI-hypersensitive sites or TFBS. The functional role of WSH at these regulatory elements remains to be investigated. We applied the WSH scores to an Ewing sarcoma dataset and showed that qFDRP can be used to reliably estimate tumor purity. It remains to be shown whether qFDRP can also be used as a reliable predictor of tumor purity for other cancer types. We created the R-package *WSHPackage* implementing all WSH scores for routine integration into existing analysis workflows. In the future, qFDRP can be used for DNA methylation-based segmentation of the genome into highly heterogeneous regions (HHRs) and lowly heterogeneous regions (LHRs) using a hidden-markov model similar to *MethylSeekR* [153]. We envision that WSH scores will contribute to the revelation of novel regulatory regions in the genome, which cannot be identified by the average DNA methylation level.

Throughout this thesis, we presented different software tools for addressing heterogeneity in DNA methylation data. Notably, the tools benefit from each other. For instance, the deconvolution protocol uses *RnBeads* as an integral part for processing DNA methylation data. Similarly, *RnBeads* is used as the DNA methylation processing tool for *MAGAR* and for storing DNA methylation data and metadata for the *WSHPackage*. We envision that further integration between the tools can jointly illuminate different levels of heterogeneity. For instance, *MAGAR* could be integrated with the deconvolution pipeline to determine tissue-specific methQTLs from bulk DNA methylation data. Additionally, deconvolution can be used in combination with WSH scores, since different proportions of cellular components across different samples can be estimated using deconvolution analysis, which is valuable information for the WSH scores. Lastly, qFDRP can be used to stratify the genome into highly and lowly variably methylated regions, which could be useful for feature selection within *DecompPipeline*.

In addition to the software tools that we presented, we also took a substantial step forward toward biologically interpreting the results generated by the different tools. For instance, we found indications of differential immune infiltration in lung adenocarcinoma and melanoma data and associations of LMCs with patient survival. Additionally, tissue-specific methQTLs have not been investigated to date and we identified both common and tissue-specific methQTLs with distinct biological properties. This will be valuable information for investigating genetic and epigenetic regulation in the context of diseases. Lastly, we provided a new estimate of tumor purity through qFDRP, which is yet to be validated in independent datasets and different cancer types. We revealed putative regulatory regions using qFDRP, which remain to be further characterized.

6.2 Outlook

The importance of DNA methylation as a clinical biomarker for various diseases will further increase, since it has major advantages in comparison to other types of molecular data. These include robustness regarding environmental influences, the quantitative readout restricted to the $[0, 1]$ interval, and the standardization of DNA methylation profiling through the Illumina microarray series. In the future, integration of multiple types of molecular data will become important due to the increasing number of publicly available datasets. On the one hand, technologies are steadily improving and applicable for mapping different epigenetic layers, such as chromatin accessibility and DNA methylation, at once. On the other hand, novel computational methods perform data integration across different types of molecular data (Figure 6.1).

Single-cell bisulfite sequencing suffers from high dropout rate, low genomic coverage, and high sequencing costs, which result in extremely sparse data matrices [300]. Thus, new technologies are required to make single-cell bisulfite sequencing broadly available to the scientific community, for instance using enzymatic treatment instead of bisulfite conversion [301]. The number of single-cell datasets steadily increases, but is still far from reaching sample sizes required to perform EWAS. It will be necessary to use datasets generated on bulk samples in combination with single-cell datasets. These could be used for instance as spiked-in reference profiles in deconvolution analysis. By these means, only a small number of samples needs to be analyzed using single-cell bisulfite sequencing, while the majority of samples is analyzed using bulk DNA methylation profiling. Since DNA methylation is particularly stable over multiple rounds of cellular divisions (neglecting DNA methylation erosion), DNA methylation may function as an ‘epigenetic memory’ of the cell. Using this notion, DNA methylation is a premier candidate for tracking the cell-of-origin of an aberration, which is especially relevant in cancer studies [302]. Single-cell bisulfite sequencing datasets will become a valuable resource to investigate cellular development over time and may be especially useful to illuminate stem cell heterogeneity. The software tools that we presented throughout this work facilitate the integration of single-cell DNA methylation with existing bulk profiles.

Third-generation sequencing technologies, including Oxford Nanopore and PacBio sequencing, yield substantially longer reads than Illumina sequencing (up to 10-100 kb), which is especially useful to resolve methylation haplotypes and cell-type-specific methylation profiles. In turn, these technologies facilitate the characterization of the relationship between genotype and DNA methylation states on the haplotype level and improve the investigation of WSH in biological samples. Finally, the combination of long-read and single-cell sequencing will allow for the quantification of cell-type and haplotype-specific methylation profiles. New technologies need to be supplemented with novel software solutions for data analysis and integration across the different data layers investigated (Figure 6.1). Understanding the biological implications of generated results is similarly important, since software needs to guide non-expert users toward data interpretation. It is critical that bioinformaticians collaborate with biologists to understand which computational solutions are currently missing in the community.

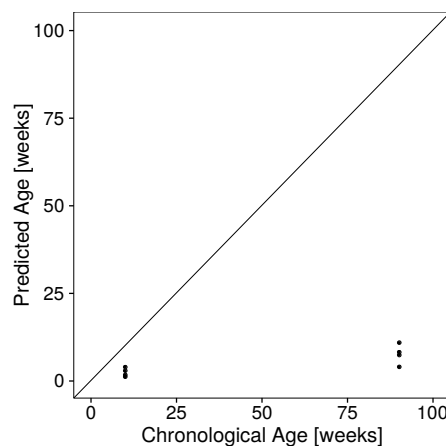
Epigenetic regulation is a complex mechanism that involves between-group, between-sample, and within-sample heterogeneity. So far, most epigenomic studies addressed individual levels of heterogeneity separately. Throughout this work, we showed that each level is complex in itself and presented novel software tools for addressing heterogeneity in DNA methylation data. We showed that the software solutions are critical for obtaining biological insights and found

that the methods are especially suitable for investigating tumor heterogeneity. We envision that the presented software packages, along with technological developments, will contribute to further our understanding of epigenetic gene regulation.

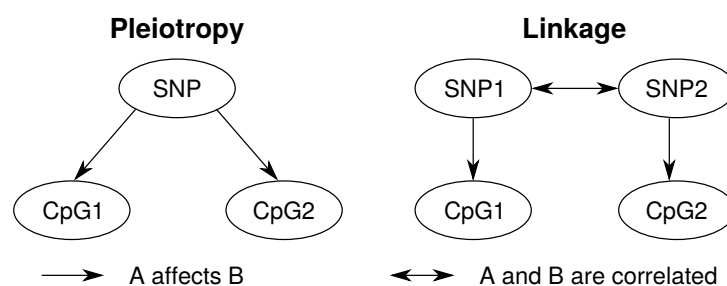
Appendix

A.1 Supplementary Figures

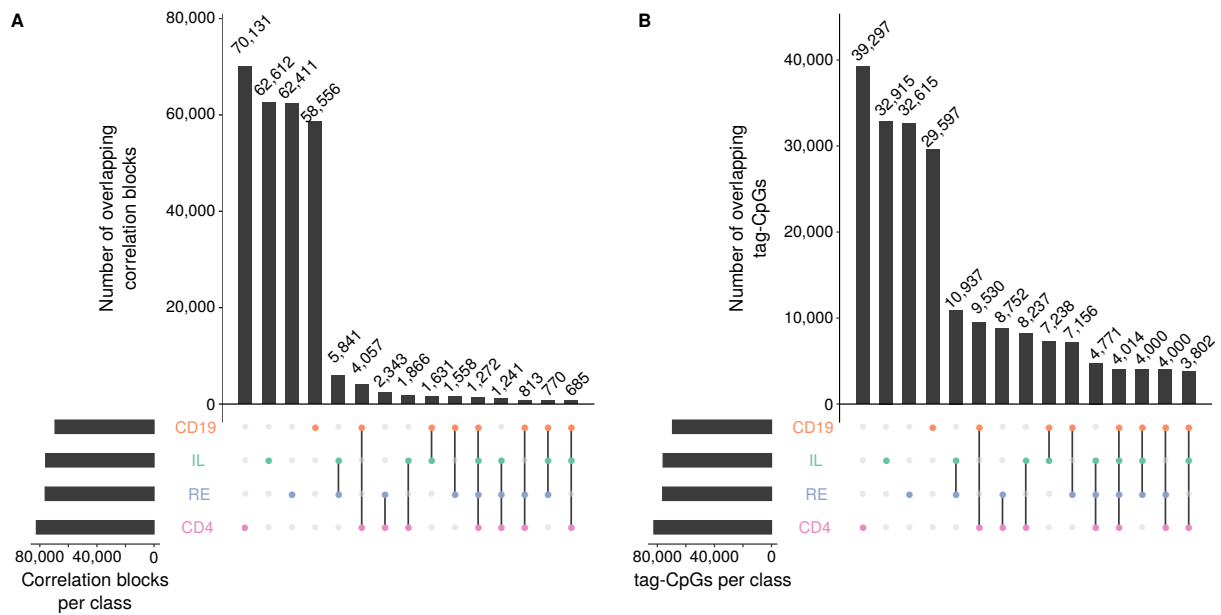
A.1.1 Chapter 3: DNA Methylation Heterogeneity Between Phenotypes



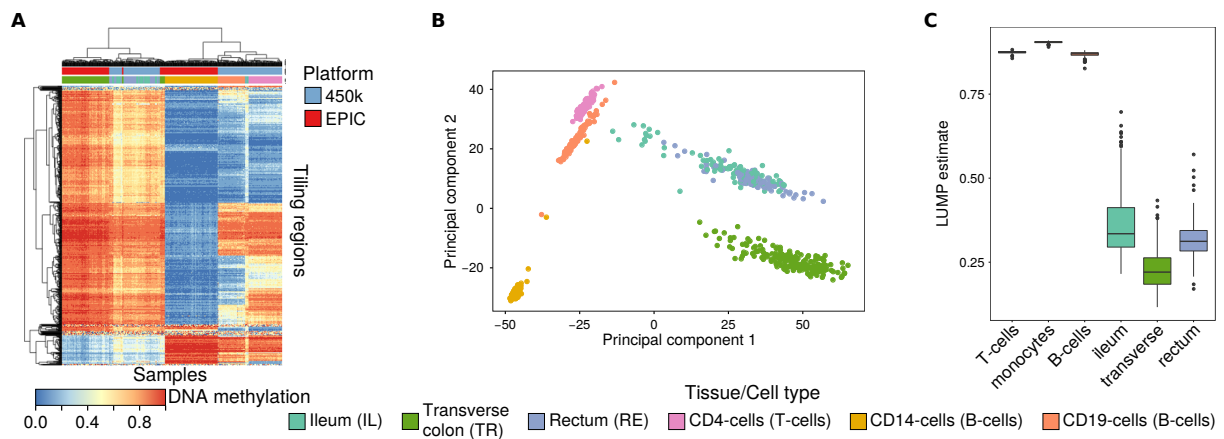
Supplementary Figure A.1: Scatterplot describing the predicted, epigenetic age (y-axis) and the annotated, chronological age (x-axis) for eight murine liver samples. Epigenetic age prediction was conducted using the murine epigenetic age predictor from Stubbs et al. [178].



Supplementary Figure A.2: Difference between pleiotropy and linkage. Pleiotropy (left) is the observation that the same SNP affects two traits (here CpG methylation states). In contrast, linkage (right) relates to two SNP that independently influence two CpGs, but that are highly correlated.

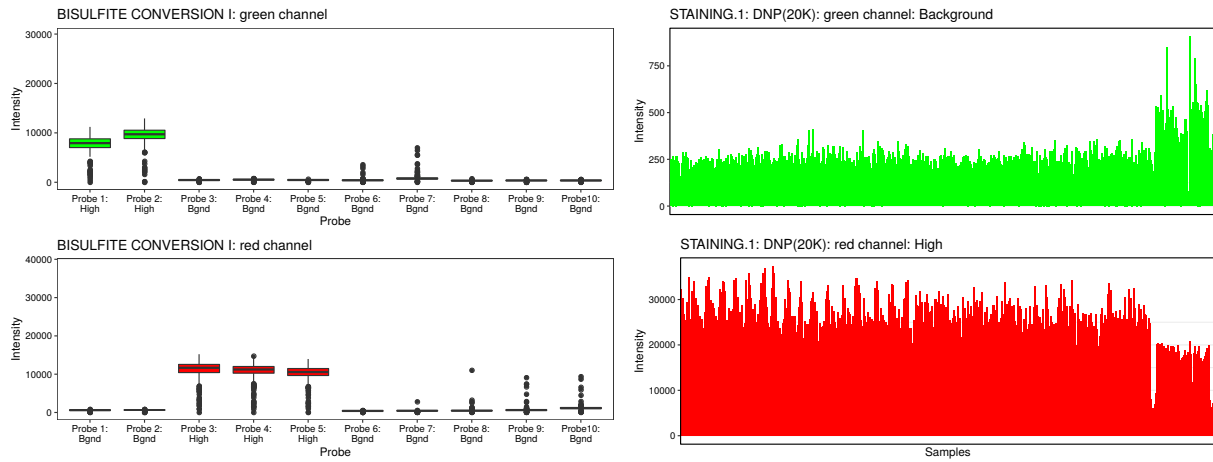


Supplementary Figure A.3: Overlapping CpG correlation blocks (**A**) and tag-CpGs (**B**) per cell type/tissue using the default parameter setting. **A:** Overlapping CpG correlation blocks for the four tissues/cell types assayed on the EPIC array. Correlation blocks were considered identical if all CpGs in the two correlation blocks defined in the two tissues independently were shared. **B:** Overlapping tag-CpGs for each of the correlation blocks per tissue/cell type. Tag-CpGs were computed for each of the correlation blocks and each of the tissues/cell types independently.

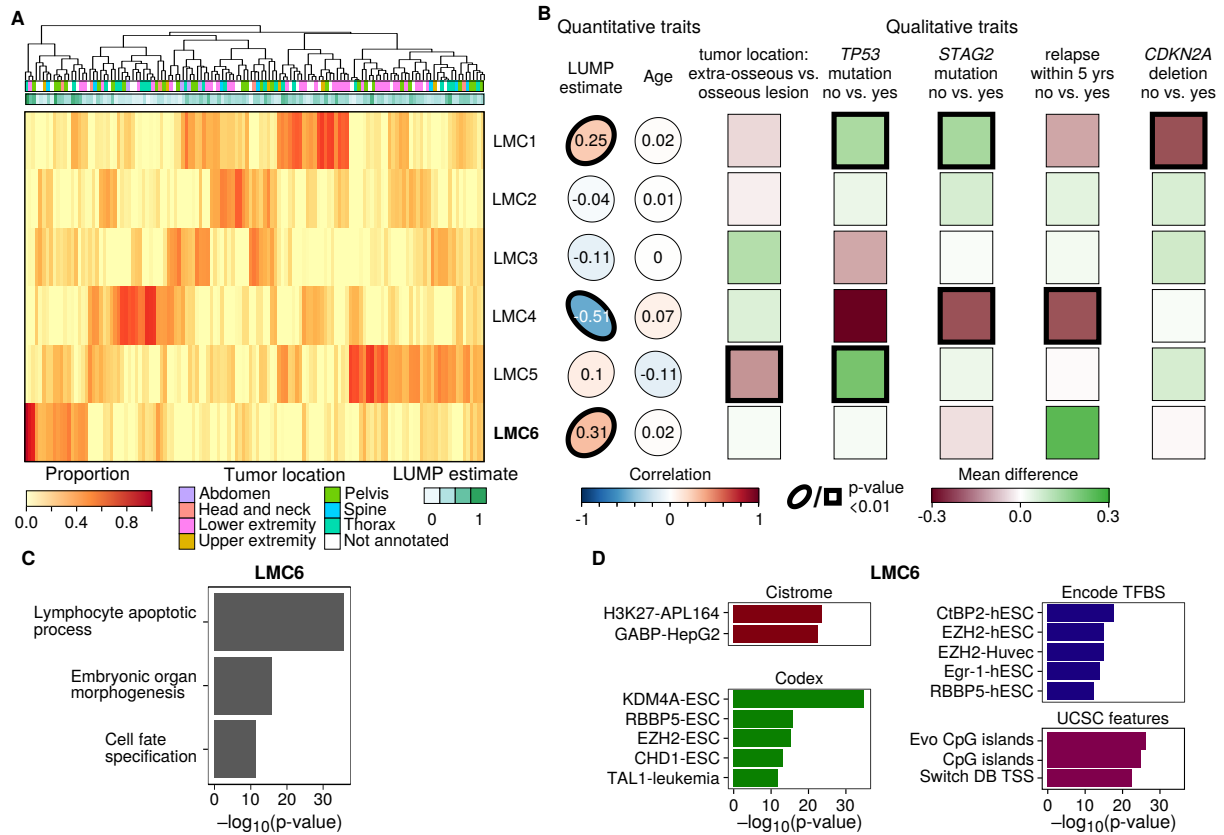


Supplementary Figure A.4: Joint description of the validation and discovery dataset for methQTL analysis (CEDAR cohort). **A:** Heatmap for methylation states of genome-wide tiling regions (5 kb) for a combination of the discovery data set (EPIC) and the validation data set (450k) from the CEDAR cohort. Analysis was restricted to the intersection between the 450k and EPIC CpGs. **B:** PCA plot for CpG-wise DNA methylation beta-values for the different samples in the combined EPIC/450k dataset. **C:** LUMP estimates of the overall immune cell content stratified according to the six different tissues/cell types.

A.1.2 Chapter 4: DNA Methylation Heterogeneity Between Samples Sharing a Phenotype

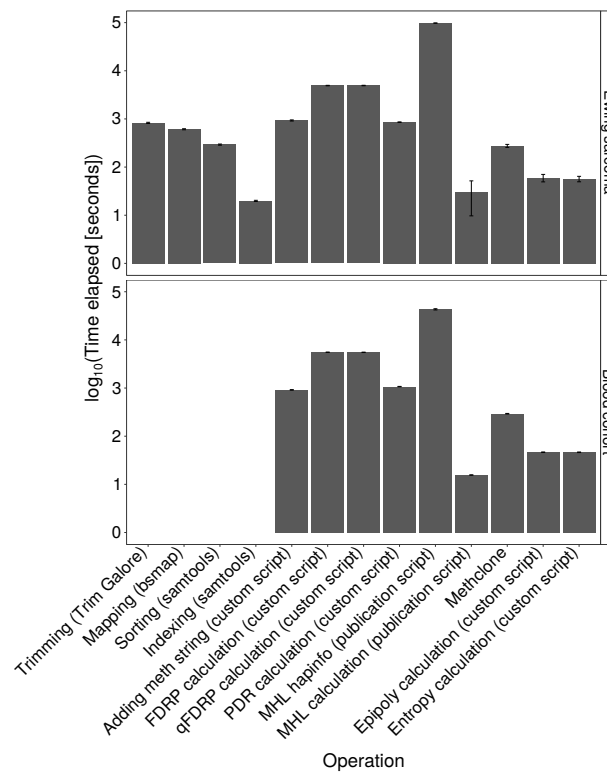


Supplementary Figure A.5: Quality control for the new samples of the CEDAR cohort. Quality control box- and barplots for bisulfite conversion and staining control probes for the red and green channels, respectively. The batch of samples shown here was expected to show low technical quality due to potential contamination of the input material. Shown is substantially lower than expected signal intensity for the high bisulfite conversion control probes and high signal intensities for the background control probes. The samples have been re-analyzed and exhibited better technical quality. This figure should serve as an example of bad-quality microarray data.

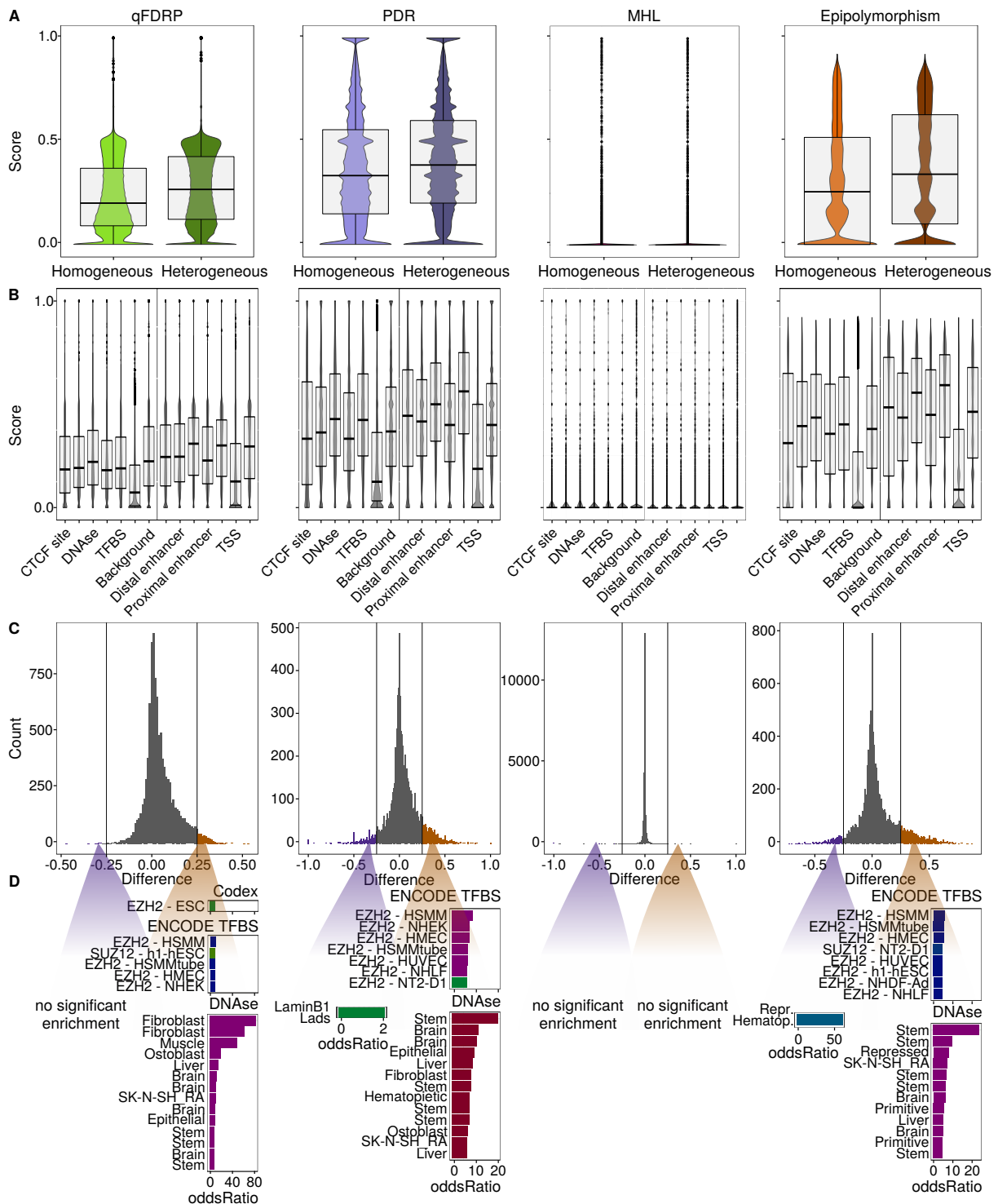


Supplementary Figure A.6: Interpreting *MeDeCom* results on the Ewing sarcoma RRBS dataset with *FactorViz*. **A:** Heatmap of LMC proportions in the Ewing sarcoma samples ($k=6$ LMCs, $\lambda=0.001$). The samples were hierarchically clustered according to the Euclidean distance between the proportions using complete linkage. We annotated samples using the tumor location and with the sample-specific LUMP estimate. **B:** Associations between the phenotypic traits and proportions. For quantitative traits, the Pearson correlations are shown as ellipses that are directed to the upper right for positive and to the lower right for negative correlations, respectively. For qualitative traits, the absolute difference of the proportions in the two groups (e.g., mutation vs. wildtype) is shown. P-values (two-sided correlation test for quantitative and two-sided t-test for categorical variables) less than 0.01 are indicated by bold outlines. GO (**C**) and LOLA (**D**) enrichment analysis of the LMC6-specific hypomethylated sites. No significant LOLA and GO enrichments were found for the remaining LMCs. Sites were defined as LMC-specific hypomethylated if the difference between the value of the LMC and the median of all other components was less than 0.5. P-values have been adjusted for multiple testing with the Benjamini-Hochberg method.

A.1.3 Chapter 5: DNA Methylation Heterogeneity Within Samples



Supplementary Figure A.7: Elapsed time for the computation of WSH scores. Histograms of wall clock time used for the different computation steps for the two RRBS data sets Ewing sarcoma (upper) and the blood cohort (lower). Shown is the common logarithm of elapsed time for the computations in seconds for one sample and the standard deviation across the samples. For the blood cohort, sequencing reads were already aligned to a reference genome using *bmap* during a preprocessing step. Thus, the first four steps were omitted.



Supplementary Figure A.8: WSH scores in homogeneous and heterogeneous hybrid samples comprising similar or distinct samples from DEEP. **A:** Genome-wide distribution of WSH scores. **B:** Stratification of genomic locations according to the Ensembl Regulatory Build and distributions of WSH scores within each of the locations defined. **C:** Histogram of WSH score differences aggregated over promoters between the heterogeneous and homogeneous sample. We set 0.25 as threshold and performed LOLA enrichment analysis (**D**) of the promoters with higher heterogeneity in the homogeneous (purple) or heterogeneous (brown) sample.

A.2 Supplementary Tables

Supplementary Table A.1 presents different tools for the analysis of DNA methylation data. We selected the most widely-used tools according to literature research and according to the Bioconductor download statistics. The table visualizes the features that are available in the tools and classifies the features according to the *RnBeads* modules *Input*, *Preprocessing*, *Quality control*, *Phenotype/covariate inference*, *Data exploration*, *Differential analysis*, and *Interface*.

Supplementary Table A.1: Feature comparison table of different software packages for the analysis of DNA methylation data.

Software	Summary	Platform			Input				Preprocessing			
		450k	EPIC	Bisulfite-seq	IDAT files	Beta value table	Read count table	Other / remarks	Filtering	Normali- zation	Smoothing	Other / remarks
ARRmNormalization	R package for 450k data normalization	✓						Probe intensity matrices		✓		
BEAT	R package for modeling of DNA methylation levels in low-input and single-cell bisulfite sequencing datasets			✓			✓					Statistical methylation modeling
BEclear	R package for batch effect correction	✓	✓	✓		✓						Missing value imputation
BiSeq	R package for differential analysis of bisulfite sequencing data			✓			✓	Bismark output		✓	✓	
bsseq/BSmooth	R package for smoothing of bisulfite-sequencing data and identification of DMRs			✓			✓			✓	✓	
ChAMP	R package for processing microarray data, including filtering, normalization, batch effect correction, DMR, and CNV calling	✓	✓		✓				✓	✓		Normalization: BMIQ, SWAN, PBC, Funnorm
COHCAP	R package specializing in the analysis of CpG islands	✓	✓	✓			✓	Bismark output	✓	✓		
conumee	R package for detecting CNVs from 450k data	✓	✓					via minfi				
DMRcaller	R package for differential methylation analysis for CpGs and non-CpGs			✓				Bismark output				
DMRcate	R package for the de novo identification of DMRs	✓	✓	✓				R data objects				
DSS	R package for identifying differential methylation (and expression) from sequencing experiments			✓			✓					
edgeR	R package for differential analysis of gene expression and methylation data			✓			✓			✓		Dispersion and size factor normalization
ENmix	R package for quality control and preprocessing of methylation array datasets	✓	✓		✓				✓	✓		Signal background correction; dye-bias and probe-type adjustment; batch-effect correction
EpiDISH	R package for the reference-based estimation of cell-type heterogeneity	✓	✓			✓						
FaST-LMM-EWASher	Command-line tool for addressing for cell-type heterogeneity in differential analysis	✓	✓			✓						
GenomeStudio	Illumina's software for analyzing methylation arrays	✓	✓		✓				✓	✓		
methyAnalysis	R package for DNA methylation data analysis and visualization	✓						R data objects (e.g. lumi)				
MethylAid	R package for interactive QC of methylation array data	✓	✓		✓							
methyKit	R package for the analysis of bisulfite sequencing experiments, including visualization, DMR detection, and batch effect correction			✓			✓	Bismark output, text files	✓			Filtering by coverage
methyIPipe	R package for CpG and non-CpG methylation from bisulfite-sequencing data			✓		✓	✓	Bismark output, BAM alignment files			✓	
MethylSeekR	R package for DNA methylation-based segmentation			✓			✓		✓			Filtering: SNPs
methyumi	R package for low-level processing of methylation arrays	✓			✓					✓		Signal background correction
metilene	Command-line tool for detection of DMCs and DMRs from bisulfite-sequencing data			✓		✓						
minfi	Various algorithms for the analysis, correction, and visualization of Infinium data	✓	✓		✓					✓		Normalization: Illumina, SWAN, Quantile, Noob, Funnorm
missMethyl	R package for analyzing methylation array data	✓	✓					via minfi		✓		Normalization: SWAN
ReffreeEWAS	R package for the reference-free estimation of intra-sample heterogeneity	✓	✓			✓						
RnBeads	Comprehensive analysis of DNA methylation data for microarrays and bisulfite sequencing	✓	✓	✓	✓	✓	✓	Parsers for Bismark, BisSNP, and various other standard formats	✓	✓		Various filters, Signal background correction, Normalization: BMIQ, SWAN, Funnorm, Missing value imputation
shinyMethyl	R shiny tool extending minfi for visualizing microarray data	✓	✓					via minfi				via minfi
wateRmelon	R package implementing a broad range of Infinium quality metrics and normalization methods	✓	✓		✓			GenomeStudio reports	✓	✓		Filtering: detection p-value, SNPs, Signal background correction, Normalization: SWAN, BMIQ

Feature comparison table of different software packages for the analysis of DNA methylation data (continued).

Software	Quality control			Phenotype / covariate inference				Data exploration					
	Control probes	Read coverage	Other / remarks	Batch-effects	CT heterogeneity	Age prediction	Other / remarks	Region-based analysis	Dim. reduction	Cluster analysis	CNVs	Genome-browser tracks	Other / remarks
ARRmNormalization													
BEAT													
BEclear				✓			Batch-effect correction (latent factor models)						
BiSeq		✓											
bsseq/BSmooth													
ChAMP			Array probe-type QC	✓	✓		ComBat, Reference-based estimation of within-sample heterogeneity (Houseman approach)		✓	✓	✓	✓	PCA/SVD, Methylation variable probes
COHCAP								✓	✓	✓			Region-based analysis for CpG islands, Methylation value distribution, Correlation to gene expression data
conumee											✓		
DMRcaller		✓										✓	
DMRcate												✓	Genome-browser plots of DMRs
DSS													
edgeR													
ENmix	✓												
EpiDISH					✓		Reference-based estimation of within-sample heterogeneity						
FaST-LMM-EWASher				✓	✓		Reference-free estimation of within-sample heterogeneity (linear mixed models)						
GenomeStudio	✓									✓			Methylation value distribution, Correlation to gene expression data
methyAnalysis												✓	Correlation of proximal CpG methylation levels
MethylAid	✓		Probe detection p-value QC										
methyKit									✓				Region-based analysis for tiling regions, Sample correlations
methyIPipe								✓					
MethylSeekR								✓					Methylation-based segmentation
methyLumi	✓		Probe detection p-value QC										
metilene													
minfi	✓	✓			✓	✓	Reference-based estimation of within-sample heterogeneity (Houseman)		✓				Methylation-based segmentation
missMethyl													
RefFreeEWAS				✓	✓		Reference-free estimation of within-sample heterogeneity (SVD)						
RnBeads	✓	✓	Multiple SNP-based QC plots	✓	✓	✓	SVA; Sex prediction; Immune cell content; Reference-based and reference-free estimation of within-sample heterogeneity (Houseman, FaST-LMM-EWASher, RefFreeEWAS)	✓	✓	✓	✓	✓	Methylation-based segmentation
shinyMethyl	✓		Array probe-type QC				Sex prediction		✓				Visualization of covariates
wateRmelon			Various QC metrics: Imprinting, X inactivation, Array probe-type, SNP-based metrics, Outlier detection		✓	✓	Reference-based estimation of within-sample heterogeneity (Houseman)						

Feature comparison table of different software packages for the analysis of DNA methylation data (continued).

[illegible]

Supplementary Table A.2: Common methQTLs identified using colocalization analysis. Beta=slope of methQTL, SE=standard error of slope, Dist=distance, p-adj=FDR-adjusted p-value

Tissue	CpG	SNP	Beta	SE	p-value	Chr	Pos (CpG)	Pos (SNP)	Dist.	p-adj
CD19	cg04811114	rs12122827	0.143	0.01	2.41E-22	chr1	202172778	202172769	9	8.24E-16
CD19	cg21718113	rs4085613	0.183	0.01	1.54E-28	chr1	152571909	152550018	21891	1.33E-21
CD19	cg01680303	rs2513559	0.187	0.01	5.05E-29	chr11	87776411	87744730	31681	3.06E-22
CD19	cg02147364	rs57144794	0.21	0.01	2.47E-26	chr12	9966812	9950728	16084	3.73E-19
CD19	cg09477447	rs8109401	-0.113	0	1.60E-38	chr19	27733670	27739429	-5759	1.74E-32
CD19	cg13593090	rs7250843	-0.133	0.01	3.19E-18	chr19	9546723	9427235	119488	1.47E-12
CD19	cg27634071	rs2272804	0.066	0.01	3.46E-12	chr22	45809740	45809624	116	1.28E-06
CD19	cg04553112	rs11927101	0.187	0.01	6.21E-28	chr3	125709451	125697489	11962	1.54E-20
CD19	cg01394167	rs12501535	0.145	0.01	2.75E-22	chr4	9479622	9456977	22645	1.60E-15
CD19	cg09255157	rs67822595	-0.232	0.01	2.94E-45	chr4	106553472	106556961	-3489	1.70E-37
CD19	cg15259449	rs10021193	-0.027	0	3.35E-15	chr4	130959885	130959880	5	5.98E-09
CD19	cg16201418	rs10024983	-0.078	0.01	3.81E-17	chr4	61367032	61367045	-13	8.31E-11
CD19	cg13944838	rs55901738	-0.201	0.01	6.60E-28	chr5	179740914	179741374	-460	2.83E-21
CD19	cg01874867	rs705379	0.112	0.01	1.47E-21	chr7	94954059	94953895	164	1.32E-14
CD19	cg08408040	rs10244924	0.173	0.01	5.90E-40	chr7	11381133	11379371	1762	3.67E-31
CD19	cg27294909	rs12379215	-0.048	0.01	1.30E-14	chr9	77555577	77555710	-133	1.96E-08
CD4	cg05044291	rs12122827	0.129	0.01	3.62E-43	chr1	202172867	202172769	98	1.27E-36
CD4	cg21718113	rs4085613	0.127	0.01	6.62E-30	chr1	152571909	152550018	21891	9.82E-24
CD4	cg01680303	rs2513559	0.151	0.01	2.04E-50	chr11	87776411	87744730	31681	3.12E-43
CD4	cg02147364	rs57144794	0.179	0	1.17E-69	chr12	9966812	9950728	16084	9.65E-62
CD4	cg09477447	rs8109401	-0.157	0.01	8.54E-48	chr19	27733670	27739429	-5759	9.69E-42
CD4	cg15727925	rs7250843	-0.144	0.01	7.15E-45	chr19	9546735	9427235	119500	7.56E-39
CD4	cg22884516	rs2272804	0.024	0	1.20E-17	chr22	45809543	45809624	-81	1.86E-12
CD4	cg04553112	rs11927101	0.185	0.01	1.02E-44	chr3	125709451	125697489	11962	4.76E-37
CD4	cg00598449	rs67822595	-0.026	0	2.05E-15	chr4	106553832	106556961	-3129	9.05E-10
CD4	cg01394167	rs12501535	0.171	0.01	8.64E-43	chr4	9479622	9456977	22645	7.00E-36
CD4	cg15259449	rs10021193	-0.033	0	2.09E-15	chr4	130959885	130959880	5	9.20E-10
CD4	cg16201418	rs10024983	-0.103	0	4.02E-40	chr4	61367032	61367045	-13	2.28E-33
CD4	cg23248424	rs55901738	-0.213	0.01	1.36E-52	chr5	179741104	179741374	-270	5.46E-45
CD4	cg08408040	rs10244924	0.157	0	3.08E-64	chr7	11381133	11379371	1762	1.53E-55
CD4	cg19678392	rs705379	0.122	0	8.47E-46	chr7	94953810	94953895	-85	1.92E-38
CD4	cg02555883	rs12379215	-0.063	0	1.51E-29	chr9	77555655	77555710	-55	6.73E-23
IL	cg21718113	rs4085613	0.153	0.01	9.01E-29	chr1	152571909	152550018	21891	4.15E-22
IL	cg26347746	rs12122827	0.162	0.01	7.29E-25	chr1	202172848	202172769	79	2.57E-18
IL	cg01680303	rs2513559	0.07	0.01	2.40E-18	chr11	87776411	87744730	31681	8.34E-12
IL	cg02147364	rs57144794	0.173	0.01	2.52E-41	chr12	9966812	9950728	16084	8.59E-34
IL	cg09477447	rs8109401	-0.107	0	5.11E-50	chr19	27733670	27739429	-5759	5.22E-44
IL	cg13593090	rs7250843	-0.129	0.01	3.23E-22	chr19	9546723	9427235	119488	2.12E-16
IL	cg22884516	rs2272804	0.04	0	3.36E-13	chr22	45809543	45809624	-81	5.66E-08
IL	cg04553112	rs11927101	0.246	0.01	5.05E-33	chr3	125709451	125697489	11962	5.14E-26
IL	cg00598449	rs67822595	-0.148	0	5.47E-47	chr4	106553832	106556961	-3129	5.79E-39
IL	cg12006118	rs12501535	0.078	0.01	1.25E-16	chr4	9479947	9456977	22970	2.19E-10
IL	cg15259449	rs10021193	-0.033	0	5.26E-22	chr4	130959885	130959880	5	1.14E-15
IL	cg16201418	rs10024983	-0.084	0.01	1.12E-25	chr4	61367032	61367045	-13	7.80E-19
IL	cg23248424	rs55901738	-0.191	0.01	9.82E-37	chr5	179741104	179741374	-270	2.52E-29
IL	cg08408040	rs10244924	0.126	0.01	1.98E-29	chr7	11381133	11379371	1762	4.24E-22
IL	cg17330251	rs705379	0.164	0.01	5.50E-28	chr7	94953956	94953895	61	1.09E-20
IL	cg02555883	rs12379215	-0.109	0.01	3.88E-20	chr9	77555655	77555710	-55	2.39E-13
RE	cg12650227	rs4085613	0.232	0.02	1.88E-21	chr1	152572930	152550018	22912	1.47E-14
RE	cg26347746	rs12122827	0.127	0.01	7.34E-22	chr1	202172848	202172769	79	1.29E-14
RE	cg01680303	rs2513559	0.057	0.01	2.56E-12	chr11	87776411	87744730	31681	9.85E-07
RE	cg02147364	rs57144794	0.188	0.01	1.09E-30	chr12	9966812	9950728	16084	2.40E-23
RE	cg09477447	rs8109401	-0.102	0.01	5.29E-25	chr19	27733670	27739429	-5759	3.85E-19
RE	cg15727925	rs7250843	-0.106	0.01	6.27E-20	chr19	9546735	9427235	119500	3.47E-14
RE	cg27634071	rs2272804	0.049	0	2.44E-17	chr22	45809740	45809624	116	2.94E-12
RE	cg04553112	rs11927101	0.264	0.02	1.06E-24	chr3	125709451	125697489	11962	7.77E-18
RE	cg00598449	rs67822595	-0.151	0.01	1.27E-31	chr4	106553832	106556961	-3129	2.40E-24
RE	cg01394167	rs12501535	0.171	0.01	1.45E-23	chr4	9479622	9456977	22645	1.09E-16
RE	cg15259449	rs10021193	-0.038	0	2.13E-20	chr4	130959885	130959880	5	1.02E-13
RE	cg16201418	rs10024983	-0.084	0.01	1.15E-22	chr4	61367032	61367045	-13	6.74E-16
RE	cg23248424	rs55901738	-0.216	0.01	5.76E-32	chr5	179741104	179741374	-270	2.06E-24
RE	cg08408040	rs10244924	0.119	0.01	8.54E-19	chr7	11381133	11379371	1762	3.59E-12
RE	cg17330251	rs705379	0.157	0.01	1.54E-28	chr7	94953956	94953895	61	4.55E-21
RE	cg02555883	rs12379215	-0.115	0.01	1.21E-21	chr9	77555655	77555710	-55	3.37E-14

A.3 Abbreviations

<i>27k-array</i>	Illumina Infinium HumanMethylation27 BeadChip
<i>450k-array</i>	Illumina Infinium HumanMethylation450 BeadChip
<i>5caC</i>	5-carboxylcytosine
<i>5fC</i>	5-formylcytosine
<i>5hmC</i>	5-hydroxymethylcytosine
<i>5mc</i>	5-methylcytosine
<i>AML</i>	Acute Myeloid Leukemia
<i>ATAC-seq</i>	Assay for Transposase-Accessible Chromatin using Sequencing
<i>AUC</i>	Area Under the Curve
<i>BAM</i>	Binary Alignment Format
<i>BED</i>	Browser Extensible Data
<i>BER</i>	Base Excision Repair
<i>BMI</i>	Body Mass Index
<i>BMIQ</i>	Beta-Mixture Quantile
<i>bp</i>	base pair
<i>caQTL</i>	Chromatin Accessibility QTL
<i>CGI</i>	CpG Island
<i>ChIP-seq</i>	Chromatin Immunoprecipitation Sequencing
<i>CIMP</i>	CGI Methylator Phenotype
<i>CLL</i>	Chronic Lymphocytic Leukemia
<i>CNV</i>	Copy Number Variation
<i>CpG</i>	Cytosine-Guanine dinucleotide
<i>CPM</i>	Counts Per Million
<i>CTCF</i>	CCCTC-Binding Factor
<i>CV</i>	Cross-Validation
<i>de.NBI</i>	German Network for Bioinformatics Infrastructure
<i>DMC</i>	Differentially Methylated Cytosine
<i>DMR</i>	Differentially Methylated Region
<i>DNA</i>	Deoxyribonucleic Acid
<i>DNAseI-seq</i>	DNAseI-hypersensitive Sites Sequencing
<i>DNMT</i>	DNA Methyltransferase
<i>dNTP</i>	deoxyribose Nucleoside Tri-Phosphate
<i>EGA</i>	European Genome-phenome Archive
<i>EPIC-array</i>	Illumina Infinium MethylationEPIC BeadChip
<i>eQTL</i>	expression QTL
<i>ESC</i>	Embryonic Stem Cell
<i>EWAS</i>	Epigenome-Wide Association Study
<i>FDR</i>	False Discovery Rate
<i>FDRP</i>	Fraction of Discordant Read Pairs
<i>FFPE</i>	Formalin-Fixed and Paraffin-Embedded
<i>FN</i>	False Negative

<i>FP</i>	False Positive
<i>GEO</i>	Gene Expression Omnibus
<i>GO</i>	Gene Ontology
<i>GUI</i>	Graphical User Interface
<i>GWAS</i>	Genome-Wide Association Study
<i>HEIDI</i>	Heterogeneity in Dependent Instruments
<i>HMD</i>	Highly Methylated Domain
<i>HMM</i>	Hidden-Markov Model
<i>HPC</i>	High Performance Computing
<i>HTML</i>	Hypertext Markup Language
<i>ICA</i>	Independent Component Analysis
<i>ICI</i>	Immune Checkpoint Inhibition
<i>IDAT</i>	Intensity Data
<i>IHEC</i>	International Human Epigenome Consortium
<i>iPSC</i>	induced Pluripotent Stem Cell
<i>kb</i>	kilobase
<i>KNN</i>	K-Nearest Neighbors
<i>LMC</i>	Latent Methylation Component
<i>LMR</i>	Lowly Methylated Region
<i>MAF</i>	Minor Allele Frequency
<i>mb</i>	megabase
<i>MBD</i>	Methyl-CpG-Binding Domain
<i>MeDIP-seq</i>	Methylated DNA Immunoprecipitation Sequencing
<i>methQTL</i>	methylation QTL
<i>MHL</i>	Methylation Haplotype Load
<i>miRNAs</i>	microRNAs
<i>mQTL</i>	metabolomic QTL
<i>mRNA</i>	messenger RNA
<i>MSC</i>	Mesenchymal Stem Cell
<i>MSD</i>	Methylation Switching Domain
<i>NGS</i>	Next-Generation Sequencing
<i>NMF</i>	Non-negative Matrix Factorization
<i>NOMe-seq</i>	Nucleosome Occupancy and Methylome Sequencing
<i>PCR</i>	Polymerase Chain Reaction
<i>PDR</i>	Proportion of Discordant Reads
<i>PMD</i>	Partially Methylated Domain
<i>qFDRP</i>	quantitative FDRP
<i>QTL</i>	Quantitative Trait Loci
<i>RAM</i>	Random Access Memory
<i>RNA</i>	Ribonucleic Acid
<i>RNA-seq</i>	RNA Sequencing
<i>RNAPII</i>	RNA Polymerase II
<i>ROC</i>	Receiver Operator Characteristic
<i>RRBS</i>	Reduced-Representation Bisulfite Sequencing
<i>RSS</i>	Residual Sum of Squares

<i>SAM</i>	S-Adenosylmethionine
<i>SGE</i>	Sun Grid Engine
<i>siRNAs</i>	small interfering RNAs
<i>SLURM</i>	Simple Linux Utility for Resource Management
<i>SMR</i>	Summary-data-based Mendelian Randomization
<i>SMRT</i>	Single Molecule Real-Time
<i>SNP</i>	Single Nucleotide Polymorphism
<i>SNV</i>	Single Nucleotide Variant
<i>SVD</i>	Singular Value Decomposition
<i>TAD</i>	Topologically Associated Domain
<i>TCGA</i>	The Cancer Genome Atlas
<i>TDG</i>	Thymine DNA Glycosylase
<i>TET</i>	Ten-Eleven Translocation
<i>TF</i>	Transcription Factor
<i>THR</i>	Truly Heterogeneous Region
<i>TN</i>	True Negative
<i>TP</i>	True Positive
<i>tSNE</i>	t-distributed Stochastic Neighbor Embedding
<i>TSS</i>	Transcription Start Site
<i>UMAP</i>	Uniform Manifold Approximation and Projection
<i>UMR</i>	Un-Methylated Region
<i>VCF</i>	Variant Call Format
<i>WGBS</i>	Whole-Genome Bisulfite Sequencing
<i>WSH</i>	Within-Sample Heterogeneity

A.4 List of Publications

First Author and Co-First Author Publications (in Chronological Order)

1. Müller, F.¹, **Scherer, M.**¹, Assenov, Y.¹, Lutsik, P.¹, Walter, J., Lengauer, T., and Bock, C. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* **20**, 55 (2019).
2. **Scherer, M.**, Nebel, A., Franke, A., Walter, J., Lengauer, T., Bock, C., Müller, F., and List, M. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res.* **48**, e46 (2020).
3. **Scherer, M.**, Nazarov, P.V., Toth, R., Sahay, S., Kaoma, T., Maurer, V., Vedenev, N., Plass, C., Lengauer, T., Walter, J., and Lutsik, P. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz, *Nat. Protoc.* **15**, 3240-3263 (2020).

Contributing Author Publications

1. Handl, L., Jalali, A., **Scherer, M.**, Eggeling, R., and Pfeifer, N. Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics.* **35**, i154–i163 (2019).

2. Decamps, C., Privé, F., Bacher, R., Jost, D., Waguët, A., Houseman, E.A., Lurie, E., Lutsik, P., Milosavljevic, A., **Scherer, M.**, Blum, M.G., and Richard, M. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. *BMC Bioinf.* **21**, 16 (2020).

A.5 Author Contribution Statements

A.5.1 Chapter 3: RnBeads 2.0: Comprehensive Analysis of DNA Methylation Data

Author contributions (taken from <https://doi.org/10.1186/s13059-019-1664-9>):

“FM, MS, PL, and YA developed the software, carried out the analyses, and prepared the use cases. FM, MS, and CB wrote the manuscript with input from PL and YA. JW, TL, and CB supervised the research. All authors read and approved the final manuscript.”

A.5.2 Chapter 4: Reference-free Deconvolution, Visualization and Interpretation of complex DNA Methylation Data Using DecompPipeline, MeDeCom and FactorViz

Author contributions (taken from <https://doi.org/10.1038/s41596-020-0369-6>):

“M.S. and P.L. implemented most of the computational procedures. P.L. and N.V. previously developed, published and recently updated MeDeCom for installation on Windows. S.S, M.S. and P.L. implemented FactorViz. P.V.N. and T.K. implemented consensus ICA. M.S. performed the analysis of the example datasets, and created all figures and tables. P.V.N., R.T. and V.M. provided crucial input to the analysis and interpretation, and thoroughly tested the protocol. P.L., J.W., T.L. and C.P. jointly supervised the project. M.S. and P.L. wrote the manuscript, with contributions from all co-authors. All authors read and approved the final text.”

A.6 Copyright Information

A.6.1 Figure Reprints

Figure 1.1 (<https://doi.org/10.6084/m9.figshare.5285500.v1>), Figure 2.1 (<https://doi.org/10.6084/m9.figshare.5285488.v1>), Figure 2.2 (<https://doi.org/10.6084/m9.figshare.5057566.v1>), Figure 2.3 (<https://doi.org/10.6084/m9.figshare.5285473.v1>), as well as Figure 2.6 (<https://doi.org/10.6084/m9.figshare.5285470.v1>) were created by Fabian Müller and used in his doctoral thesis [314] and are available under the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which grants copying and redistribution in any medium. All remaining figures were either created for this

¹joint first authors

thesis by the author or were copied and potentially modified from the original publication (see the next sections for detailed copyright information).

A.6.2 Chapter 3: RnBeads 2.0: Comprehensive Analysis of DNA Methylation Data

The manuscript Müller et al. [140] has been published in *Genome Biology* under the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which grants the following permission according to the publisher (taken from <https://doi.org/10.1186/s13059-019-1664-9>):

“Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.”

A.6.3 Chapter 4: Reference-free Deconvolution, Visualization and Interpretation of complex DNA Methylation Data using DecompPipeline, MeDeCom and FactorViz

The manuscript Scherer et al. [222] was published in *Nature Protocols*, which grants rights for authors to reuse the contribution in their own thesis. The information below is taken from the *Nature Research* website, which applies also to *Nature Protocols* (<https://www.nature.com/nature-research/reprints-and-permissions/permissions-requests>).

“Ownership of copyright in original research articles remains with the Author, and provided that, when reproducing the contribution or extracts from it or from the Supplementary Information, the Author acknowledges first and reference publication in the Journal, the Author retains the following non-exclusive rights:

To reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s). (...)

Authors have the right to reuse their article’s Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution.

Authors must properly cite the published article in their thesis according to current citation standards.”

A.6.4 Chapter 5: Quantitative Comparison of Within-Sample Heterogeneity Scores for DNA Methylation Data

The manuscript Scherer et al. [268] was published in *Nucleic Acids Research* under the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), granting the following permissions according to the publisher (taken from <https://doi.org/10.1093/nar/gkaa120>):

“This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.”

References

- [1] Allis, C.D. and Jenuwein, T., The molecular hallmarks of epigenetic control, *Nat. Rev. Genet.*, **17**, 487–500 (2016).
- [2] Waddington, C.H., The Epigenotype, *Endeavour* (1942).
- [3] Waddington, C.H., *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*, Allen & Unwin (1957).
- [4] F. H. Crick, On protein synthesis, *Symp. Soc. Exp. Biol.*, **12**, 138–163 (1958).
- [5] Durek, P., et al., Epigenomic Profiling of Human CD4⁺ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development, *Immunity*, **45**, 1148–1161 (2016).
- [6] Campbell, N.A., et al., *Biologie*, Pearson, 8 Auflage (2009), ISBN 9783827372871.
- [7] Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J., Crystal structure of the nucleosome core particle at 2.8 Å resolution, *Nature*, **389**, 251–260 (1997).
- [8] Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R., The Ensembl Regulatory Build, *Genome Biol.*, **16**, 56 (2015).
- [9] Jung, I., et al., A compendium of promoter-centered long-range chromatin interactions in the human genome, *Nat. Genet.*, **51**, 1442–1449 (2019).
- [10] Schmidt, F., et al., Combining transcription factor binding affinities with open-chromatin data for accurate Gene Expr. prediction, *Nucleic Acids Res.*, **45**, 54–66 (2017).
- [11] Kim, S., Yu, N.K., and Kaang, B.K., CTCF as a multifunctional protein in genome regulation and Gene Expr., *Exp. Mol. Med.*, **47**, e166–e166 (2015).
- [12] Bártová, E., Krejčí, J., Harničarová, A., Galiová, G., and Kozubek, S., Histone Modifications and Nuclear Architecture: A Review, *Journal of Histochemistry & Cytochemistry*, **56**, 711–721 (2008).
- [13] Turner, B.M., Reading signals on the nucleosome with a new nomenclature for modified histones, *Nat. Struct. Mol. Biol.*, **12**, 110–112 (2005).
- [14] Lennartsson, A. and Ekwall, K., Histone modification patterns and epigenetic codes, *Biochim. Biophys. Acta (BBA) - General Subjects*, **1790**, 863–868 (2009).
- [15] Kouzarides, T., Chromatin Modifications and Their Function, *Cell*, **128**, 693–705 (2007).
- [16] Ernst, J. and Kellis, M., Chromatin-state discovery and genome annotation with ChromHMM, *Nat. Protoc.*, **12**, 2478–2492 (2017).
- [17] Park, P.J., ChIP-seq: Advantages and challenges of a maturing technology, *Nat. Rev. Genet.*, **10**, 669–680 (2009).
- [18] Salhab, A., et al., A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains, *Genome Biol.*, **19**, 9–11 (2018).
- [19] Song, L. and Crawford, G.E., DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells, *Cold Spring Harbor Protocols*, **2010**, pdb.prot5384 (2010).
- [20] Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J., ATAC seq: A Method for Assaying Chromatin Accessibility Genome Wide, *Current Protocols in Mol. Biol.*, **109**, 21.29.1–21.29.9 (2015).
- [21] Kelly, T.K., et al., Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules, *Genome Res.*, **22**, 2497–2506 (2012).
- [22] Lieberman-Aiden, E., et al., Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, *Science*, **326**, 289–293 (2009).
- [23] Pombo, A. and Dillon, N., Three-dimensional genome architecture: players and mechanisms, *Nat. Rev. Mol. Cell Biol.*, **16**, 245–257 (2015).
- [24] Schübeler, D., Function and information content of DNA methylation, *Nature*, **517**, 321–326 (2015).
- [25] Duret, L., Mutation Patterns in the Human Genome: More Variable Than Expected, *PLoS Biol.*, **7**, e1000028 (2009).
- [26] Rideout, W., Coetzee, G., Olumi, A., and Jones, P., 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes, *Science*, **249**, 1288–1290 (1990).
- [27] Sánchez-Romero, M.A., Cota, I., and Casadesús, J., DNA methylation in bacteria: from the methyl group to the methylome, *Curr. Opin. Microbiol.*, **25**, 9–16 (2015).

- [28] Vanyushin, B.F. and Ashapkin, V.V., DNA methylation in higher plants: Past, present and future, *Biochim. Biophys. Acta (BBA) - Gene Regulatory Mechanisms*, **1809**, 360–368 (2011).
- [29] Jones, P.A., Functions of DNA methylation: islands, start sites, gene bodies and beyond, *Nat. Rev. Genet.*, **13**, 484–492 (2012).
- [30] Saxonov, S., Berg, P., and Brutlag, D.L., A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 1412–1417 (2006).
- [31] Eckhardt, F., et al., DNA methylation profiling of human chromosomes 6, 20 and 22, *Nat. Genet.*, **38**, 1378–1385 (2006).
- [32] Cooper, D., Eukaryotic DNA methylation, *Hum. Genet.*, **64**, 315–333 (1983).
- [33] Liao, J., et al., Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic Stem Cells, *Nat. Genet.*, **47**, 469–478 (2015).
- [34] Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S., DNMT3L Stimulates the DNA Methylation Activity of Dnmt3a and Dnmt3b through a Direct Interaction, *J. Biol. Chem.*, **279**, 27816–27823 (2004).
- [35] Oh, G., et al., Epigenetic assimilation in the aging human brain, *Genome Biol.*, **17**, 76 (2016).
- [36] Skvortsova, K., et al., Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA, *Epigenetics Chromatin*, **10**, 16 (2017).
- [37] Iurlaro, M., et al., In Vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine, *Genome Biol.*, **17**, 141 (2016).
- [38] Uribe-Lewis, S., et al., 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer, *Genome Biol.*, **16**, 69 (2015).
- [39] Yoder, J.A., Walsh, C.P., and Bestor, T.H., Cytosine methylation and the ecology of intragenomic parasites, *Trends Genet.*, **13**, 335–340 (1997).
- [40] Reik, W. and Walter, J., Genomic imprinting: parental influence on the genome, *Nat. Rev. Genet.*, **2**, 21–32 (2001).
- [41] Jones, P., The DNA methylation paradox, *Trends Genet.*, **15** (1999).
- [42] Smallwood, S.A. and Kelsey, G., De novo DNA methylation: a germ cell perspective, *Trends Genet.*, **28**, 33–42 (2012).
- [43] Dean, W., Santos, F., and Reik, W., Epigenetic reprogramming in early mammalian development and following somatic nuclear transfer, *Semin. Cell Dev. Biol.*, **14**, 93–100 (2003).
- [44] Horvath, S., DNA methylation age of human tissues and cell types, *Genome Biol.*, **14**, R115 (2013).
- [45] Heyn, H., et al., Distinct DNA methylomes of newborns and centenarians, *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 10522–10527 (2012).
- [46] Hannum, G., et al., Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates, *Mol. Cell*, **49**, 359–367 (2013).
- [47] Koch, C.M. and Wagner, W., Epigenetic-aging-signature to determine age in different tissues, *Aging*, **3**, 1018–1027 (2011).
- [48] Craig Venter, J., et al., The sequence of the human genome, *Science*, **291**, 1304–1351 (2001).
- [49] Lander, S., et al., Initial sequencing and analysis of the human genome International Human Genome Sequencing Consortium* The Sanger Centre: Beijing Genomics Institute/Human Genome Center, *Nature*, **409** (2001).
- [50] Ripke, S., et al., Biological insights from 108 schizophrenia-associated genetic loci, *Nature*, **511**, 421–427 (2014).
- [51] Reynisdottir, I., et al., Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2, *American J. Hum. Genet.*, **73**, 323–335 (2003).
- [52] Orozco, G., McAllister, K., and Eyre, S., Genetics of rheumatoid arthritis: GWAS and beyond, *Open Access Rheumatology: Research and Reviews*, **3**, 31 (2011).
- [53] Momozawa, Y., et al., IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes, *Nat. Commun.*, **9**, 2427 (2018).
- [54] Pierce, B.L., et al., Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms, *Nat. Commun.*, **9**, 804 (2018).
- [55] Tehranchi, A., et al., Fine-mapping cis-

- regulatory variants in diverse human populations, *eLife*, **8** (2019).
- [56] Kraus, W.E., et al., Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis, *PLoS Genet.*, **11**, e1005553 (2015).
- [57] McRae, A.F., et al., Identification of 55,000 Replicated DNA Methylation QTL., *Sci. Rep.*, **8**, 17605 (2018).
- [58] Huan, T., et al., Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease, *Nat. Commun.*, **10**, 4267 (2019).
- [59] Hannon, E., et al., Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expr., and Complex Traits., *American J. Hum. Genet.*, **103**, 654–665 (2018).
- [60] Zhao, T., Hu, Y., Zang, T., and Wang, Y., Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes, *Front. Genet.*, **10**, 1–8 (2019).
- [61] Yu, H., Cheng, W., Zhang, X., Wang, X., and Yue, W., Integration analysis of methylation quantitative trait loci and GWAS identify three schizophrenia risk variants, *Neuropsychopharmacology*, **45**, 1179–1187 (2020).
- [62] Ivanoff, S., et al., 5-Azacytidine treatment for relapsed or refractory acute myeloid leukemia after intensive chemotherapy, *Am. J. Hematol.*, **88**, 601–605 (2013).
- [63] Philibert, R., et al., A quantitative epigenetic approach for the assessment of cigarette consumption, *Front. Psychol.*, **6**, 1–8 (2015).
- [64] Gao, X., Jia, M., Zhang, Y., Breitling, L.P., and Brenner, H., DNA methylation changes of whole Blood Cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies, *Clinical Epigenetics*, **7**, 113 (2015).
- [65] Ligthart, S., et al., Tobacco smoking is associated with DNA methylation of diabetes susceptibility genes, *Diabetologia*, **59**, 998–1006 (2016).
- [66] Wahl, S., et al., Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity, *Nature*, **541**, 81–86 (2017).
- [67] Heyn, H., Moran, S., and Esteller, M., Aberrant DNA methylation profiles in the premature aging disorders Hutchinson-Gilford Progeria and Werner syndrome, *Epigenetics*, **8**, 28–33 (2013).
- [68] Liu, Y., Aryee, M., Padyukov, L., Fallin, D., and Feinberg, A., Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis, *Nat. Biotechnol.*, **13**, 142–147 (2013).
- [69] Pidsley, R., Viana, J., Spiers, H., Hannon, E., and Mill, J., Methylomic profiling of human brain tissue supports a neurodevelopmental origin of schizophrenia, *Genome Biol.*, **15**, 483 (2014).
- [70] Azuara, D., et al., New Methylation Biomarker Panel for Early Diagnosis of Dysplasia or Cancer in High-Risk Inflammatory Bowel Disease Patients, *Inflamm. Bowel Dis.*, **24**, 2555–2564 (2018).
- [71] Huynh, J., Garg, P., Thin, T., Yoo, S., and Casaccia, P., Epigenome-wide differences in pathology-free regions of multiple-sclerosis affected brains, *Nat. Neurosci.*, **17**, 121–130 (2014).
- [72] Souren, N.Y., et al., DNA methylation signatures of monozygotic twins clinically discordant for Mult. Scler., *Nat. Commun.*, **10**, 2094 (2019).
- [73] Guintivano, J., Aryee, M., and Kaminsky, Z., A Cell Epigenotype Specific Model for the Correction of Brain Cellular Heterogeneity bias and its application to age, brain region and major depression, *Epigenetics*, **8**, 290–302 (2013).
- [74] Rakyan, V.K., et al., Identification of Type 1 Diabetes-Associated DNA Methylation Variable Positions That Precede Disease Diagnosis, *PLoS Genet.*, **7**, e1002300 (2011).
- [75] Rakyan, V.K., Down, T.A., Balding, D.J., and Beck, S., Epigenome-wide association studies for common human diseases, *Nat. Rev. Genet.*, **12**, 529–541 (2011).
- [76] Karpinski, P., Pesz, K., and Sasiadek, M.M., Pan-cancer analysis reveals presence of pronounced DNA methylation drift in CpG island methylator phenotype clusters, *Epigenomics*, **9**, 1341–1352 (2017).
- [77] Tovy, A., et al., p53 is essential for DNA methylation homeostasis in naïve embryonic

- Stem Cells, and its loss promotes clonal heterogeneity, *Genes Dev.*, **31**, 959–972 (2017).
- [78] Vidal, E., et al., A DNA methylation map of human cancer at single base-pair resolution, *Oncogene*, **36**, 5648–5657 (2017).
- [79] Aryee, M.J., et al., DNA Methylation Alterations Exhibit Intraindividual Stability and Interindividual Heterogeneity in Prostate Cancer Metastases, *Sci. Transl. Med.*, **5** (2013).
- [80] Hlady, R.A., et al., Initiation of aberrant DNA methylation patterns and heterogeneity in precancerous lesions of human hepatocellular cancer, *Epigenetics*, **12**, 215–225 (2017).
- [81] Sasca, D. and Huntly, B.J.P., Independence of epigenetic and genetic diversity in AML, *Nat. Med.*, **22**, 708–709 (2016).
- [82] Gaiti, F., et al., Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia, *Nature*, **569**, 576–580 (2019).
- [83] Klughammer, J., et al., The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space, *Nat. Med.*, **24**, 1611–1624 (2018).
- [84] Møller, M., et al., Heterogeneous patterns of DNA methylation-based field effects in histologically normal prostate tissue from cancer patients, *Sci. Rep.*, **7**, 40636 (2017).
- [85] Capper, D., et al., DNA methylation-based classification of central nervous system tumours, *Nature*, **555**, 469–474 (2018).
- [86] Jones, P.A., Issa, J.P.J., and Baylin, S., Targeting the cancer epigenome for therapy, *Nat. Rev. Genet.*, **17**, 630–641 (2016).
- [87] Susan, J., Harrison, J., Paul, C.L., and Frommer, M., High sensitivity mapping of methylated cytosines, *Nucleic Acids Res.*, **22**, 2990–2997 (1994).
- [88] Frommer, M., et al., A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands., *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 1827–1831 (1992).
- [89] Zhang, Y., et al., Model-based Analysis of ChIP-Seq (MACS), *Genome Biol.*, **9**, R137 (2008).
- [90] Bibikova, M., et al., High density DNA methylation array with single CpG site resolution, *Genomics*, **98**, 288–295 (2011).
- [91] Pidsley, R., et al., Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling, *Genome Biol.*, **17**, 208 (2016).
- [92] Dedeurwaerder, S., et al., Evaluation of the Infinium Methylation 450K technology, *Epigenomics*, **3**, 771–784 (2011).
- [93] Teschendorff, A.E., et al., A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data, *Bioinformatics*, **29**, 189–196 (2013).
- [94] Fortin, J.P., et al., Functional normalization of 450k methylation array data improves replication in large cancer studies, *Genome Biol.*, **15**, 503 (2014).
- [95] Deamer, D., Akesson, M., and Branton, D., Three decades of nanopore sequencing, *Nat. Biotechnol.*, **34**, 518–524 (2016).
- [96] Eid, J., et al., Real-Time DNA Sequencing from Single Polymerase Molecules, *Science*, **323**, 133–138 (2009).
- [97] Gries, J., et al., Bi-PROF: Bisulfite profiling of target regions using 454 GS FLX Titanium technology, *Epigenetics*, **8**, 765–771 (2013).
- [98] Lister, R., et al., Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature*, **462**, 315–322 (2009).
- [99] Adey, A. and Shendure, J., Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing, *Genome Res.*, **22**, 1139–1143 (2012).
- [100] Smallwood, S.A., et al., Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity, *Nat. Methods*, **11**, 817–820 (2014).
- [101] Angermueller, C., et al., Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity., *Nat. Methods*, **13**, 229–232 (2016).
- [102] Hu, Y., et al., Simultaneous profiling of transcriptome and DNA methylome from a single cell, *Genome Biol.*, **17**, 88 (2016).
- [103] Scott, C.A., et al., Identification of cell type-specific methylation signals in bulk whole genome bisulfite sequencing data, *Genome Biol.*, **21**, 156 (2020).
- [104] Meissner, A., et al., Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis, *Nucleic Acids Res.*, **33**, 5868–5877 (2005).

- [105] Martinez-Arguelles, D.B., Lee, S., and Papadopoulos, V., In silico analysis identifies novel restriction enzyme combinations that expand reduced representation bisulfite sequencing CpG coverage, *BMC Research Notes*, **7**, 534 (2014).
- [106] Wang, J., et al., Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing, *BMC Genomics*, **14**, 11 (2013).
- [107] Weber, M., et al., Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed Hum. Cells, *Nat. Genet.*, **37**, 853–862 (2005).
- [108] Down, T.A., et al., A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis, *Nat. Biotechnol.*, **26**, 779–785 (2008).
- [109] Serre, D., Lee, B.H., and Ting, A.H., MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome, *Nucleic Acids Res.*, **38**, 391–399 (2009).
- [110] Bock, C., et al., Quantitative comparison of genome-wide DNA methylation mapping technologies, *Nat. Biotechnol.*, **28**, 1106–1114 (2010).
- [111] Aryee, M.J., et al., Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays, *Bioinformatics*, **30**, 1363–1369 (2014).
- [112] Assenov, Y., et al., Comprehensive analysis of DNA methylation data with RnBeads, *Nat. Methods*, **11**, 1138–1140 (2014).
- [113] Sherry, S.T., et al., dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, **29**, 308–311 (2001).
- [114] Chen, Y.A., et al., Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray, *Epigenetics*, **8**, 203–209 (2013).
- [115] Teschendorff, A., et al., A beta-mixture quantile normalization method for correcting probe bias in Illumina Infinium 450k DNA methylation data, *Bioinformatics*, **29**, 189–196 (2013).
- [116] Pidsley, R., et al., A data-driven approach to preprocessing Illumina 450K methylation array data, *BMC Genomics*, **14**, 293 (2013).
- [117] Maksimovic, J., Gordon, L., and Oshlack, A., SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips, *Genome Biol.*, **13**, R44 (2012).
- [118] Triche, T.J., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., and Siegmund, K.D., Low-level processing of Illumina Infinium DNA Methylation BeadArrays, *Nucleic Acids Res.*, **41**, e90–e90 (2013).
- [119] Liu, R., Dai, Z., Yeager, M., Irizarry, R.A., and Ritchie, M.E., KRLMM: an adaptive genotype calling method for common and low frequency variants, *BMC Bioinf.*, **15**, 158 (2014).
- [120] Consortium, T.I.H., A second generation human haplotype map of over 3.1 million SNPs, *Nature*, **449**, 851–861 (2007).
- [121] The 1000 Genomes Project Consortium, A global reference for human genetic variation, *Nature*, **526**, 68–74 (2015).
- [122] Howie, B.N., Donnelly, P., and Marchini, J., A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies, *PLoS Genet.*, **5**, e1000529 (2009).
- [123] Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R., MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes, *Genet. Epidemiol.*, **34**, 816–834 (2010).
- [124] Das, S., et al., Next-generation genotype imputation service and methods, *Nat. Genet.*, **48**, 1284–1287 (2016).
- [125] Stunnenberg, H.G., et al., The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery, *Cell*, **167**, 1145–1149 (2016).
- [126] Adams, D., et al., BLUEPRINT to decode the epigenetic signature written in blood, *Nat. Biotechnol.*, **30**, 224–226 (2012).
- [127] Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, **17**, 10 (2011).
- [128] Xi, Y. and Li, W., BSMAP: whole genome bisulfite sequence MAPping program, *BMC Bioinf.*, **10**, 232 (2009).
- [129] Krueger, F. and Andrews, S.R., Bismark: A flexible aligner and methylation caller for

- Bisulfite-Seq applications, *Bioinformatics*, **27**, 1571–1572 (2011).
- [130] Merkel, A., et al., gemBS: high throughput processing for DNA methylation data from bisulfite sequencing, *Bioinformatics*, **35**, 737–742 (2019).
- [131] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L., Ultrafast and memory-efficient alignment of short DNA Seq.s to the human genome, *Genome Biol.*, **10**, R25 (2009).
- [132] James, G., Witten, D., Hastie, T., and Tibshirani, R., *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*, Springer New York, New York, NY (2013), ISBN 978-1-4614-7137-0.
- [133] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, New York, NY (2001), ISBN 978-0-387-84857-0.
- [134] Tibshirani, R., Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288 (1996).
- [135] Zou, H. and Hastie, T., Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Series B. Stat. Methodol.*, **67**, 301–320 (2005).
- [136] Friedman, J., Hastie, T., and Tibshirani, R., Regularization Paths for Generalized Linear Models via Coordinate Descent., *J. Stat. Softw.*, **33**, 1–22 (2010).
- [137] Van Der Maaten, L.J.P. and Hinton, G.E., Visualizing high-dimensional data using t-sne, *J. Mach. Learn. Res.*, **9**, 2579–2605 (2008).
- [138] McInnes, L., Healy, J., and Melville, J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv* (2018).
- [139] Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E., Fast unfolding of communities in large networks, *J. Stat. Mech: Theory Exp.*, **2008**, P10008 (2008).
- [140] Müller, F., et al., RnBeads 2.0: comprehensive analysis of DNA methylation data, *Genome Biol.*, **20**, 55 (2019).
- [141] Scherer, M., Dissecting DNA Methylation in Human Aging, MSc thesis, Saarland University (2016).
- [142] Levine, M.E., et al., An epigenetic biomarker of aging for lifespan and healthspan, *Aging (Albany NY)*, **10**, 573–591 (2018).
- [143] Michels, K.B., et al., Recommendations for the design and analysis of epigenome-wide association studies, *Nat. Methods*, **10**, 949–955 (2013).
- [144] Bock, C., Analysing and interpreting DNA methylation data, *Nat. Rev. Genet.*, **13**, 705–719 (2012).
- [145] Wreczycka, K., et al., Strategies for analyzing bisulfite sequencing data, *J. Biotechnol.*, **261**, 105–115 (2017).
- [146] Chen, D.P., Lin, Y.C., and Fann, C.S., Methods for identifying differentially methylated regions for sequence- and array-based data, *Briefings in Functional Genomics*, **15**, 485–490 (2016).
- [147] Teschendorff, A.E. and Zheng, S.C., Cell-type deconvolution in epigenome-wide association studies: A review and recommendations, *Epigenomics*, **9**, 757–768 (2017).
- [148] Horvath, S. and Levine, A., HIV-1 Infection Accelerates Age According to the Epigenetic Clock, *J. Infect. Dis.*, **212**, 1563–1573 (2015).
- [149] Aran, D., Sirota, M., and Butte, A.J., Systematic pan-cancer analysis of tumour purity, *Nat. Commun.*, **6**, 8971 (2015).
- [150] Phipson, B. and Oshlack, A., DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging, *Genome Biol.*, **15**, 465 (2014).
- [151] Teschendorff, A.E., et al., DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer, *Nat. Commun.*, **7**, 10478 (2016).
- [152] Sheffield, N.C. and Bock, C., LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor, *Bioinformatics*, **32**, 587–589 (2016).
- [153] Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B., Identification of active regulatory regions from DNA methylation data., *Nucleic Acids Res.*, **41**, e155 (2013).
- [154] Sheffield, N.C., et al., DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma, *Nat. Med.*, **23**, 386–395 (2017).
- [155] Alizadeh, A.A., et al., Toward understanding and exploiting tumor heterogeneity, *Nat. Med.*, **21**, 846–853 (2015).
- [156] Tomazou, E.M., et al., Epigenome Mapping Reveals Distinct Modes of Gene Regulation

- and Widespread Enhancer Reprogramming by the Oncogenic Fusion Protein EWS-FLI1, *Cell Reports*, **10**, 1082–1095 (2015).
- [157] Riebler, A., et al., BayMeth: Improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach, *Genome Biol.*, **15**, 0–19 (2014).
- [158] Morris, T.J. and Beck, S., Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data, *Methods*, **72**, 3–8 (2015).
- [159] Davis, S., et al., methylumi: Handle Illumina methylation data (2015).
- [160] Tian, Y., et al., ChAMP: Updated Methylation Analysis Pipeline for Illumina BeadChips, *Bioinformatics*, **33**, 3982–3984 (2017).
- [161] Akalin, A., et al., methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles, *Genome Biol.*, **13**, R87 (2012).
- [162] Johansson, Å., Enroth, S., and Gyllenstein, U., Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan, *PLoS ONE*, **8** (2013).
- [163] Schillebeeckx, M., et al., Laser capture microdissection–reduced representation bisulfite sequencing (LCM-RRBS) maps changes in DNA methylation associated with gonadectomy-induced adrenocortical neoplasia in the mouse, *Nucleic Acids Res.*, **41**, e116–e116 (2013).
- [164] Troyanskaya, O., et al., Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520–525 (2001).
- [165] Szymczak, S., et al., DNA methylation QTL analysis identifies new regulators of human longevity, *Hum. Mol. Genet.*, **29**, 1154–1167 (2020).
- [166] Reizel, Y., et al., Gender-specific postnatal demethylation and establishment of epigenetic memory, *Genes Dev.*, **29**, 923–933 (2015).
- [167] Groß, M., Identifying Partially Methylated Domains Using DNA Methylation Microarray Data, MSc thesis, Saarland University (2020).
- [168] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature*, **489**, 57–74 (2012).
- [169] Sánchez-Castillo, M., et al., CODEX: A next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities, *Nucleic Acids Res.*, **43**, D1117–D1123 (2015).
- [170] Liu, T., et al., Cistrome: an integrative platform for transcriptional regulation studies, *Genome Biol.*, **12**, R83 (2011).
- [171] Kent, W.J., et al., The Human Genome Browser at UCSC, *Genome Res.*, **12**, 996–1006 (2002).
- [172] Hansen, K.D., Langmead, B., and Irizarry, R.A., BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions, *Genome Biol.*, **13**, R83 (2012).
- [173] Chen, J., Lutsik, P., Akulenko, R., Walter, J., and Helms, V., AKSmooth: Enhancing low-coverage bisulfite sequencing data via kernel-based smoothing, *J. Bioinform. Comput. Biol.*, **12**, 1442005 (2014).
- [174] Weidner, C.I., Lin, Q., and Wagner, W., Aging of Blood can be tracked by DNA methylation changes at just three CpG sites, *Genome Biol.*, **15** (2014).
- [175] Momeni, Z. and Saniee Abadeh, M., MapReduce-Based Parallel Genetic Algorithm for CpG-Site Selection in Age Prediction, *Genes*, **10**, 969 (2019).
- [176] Horvath, S. and Raj, K., DNA methylation-based biomarkers and the epigenetic clock theory of ageing, *Nat. Rev. Genet.*, **19**, 371–384 (2018).
- [177] Marioni, R.E., et al., DNA methylation age of blood predicts all-cause mortality in later life, *Genome Biol.*, **16** (2015).
- [178] Stubbs, T.M., et al., Multi-Tissue DNA Methylation Age Predictor In Mouse, *Genome Biol.*, **19**, 68 (2017).
- [179] Han, Y., et al., Epigenetic age-predictor for mice based on three CpG sites, *eLife*, **7**, 1–10 (2018).
- [180] Frobel, J., et al., Epigenetic rejuvenation of mesenchymal stromal cells derived from induced pluripotent Stem Cells, *Stem Cell Rep.*, **3**, 414–422 (2014).
- [181] Maierhofer, A., et al., Accelerated epigenetic aging in Werner syndrome, *Aging*, **9**, 1143–1152 (2017).
- [182] Marioni, R.E., et al., The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936, *Inter-*

- Natl. J. (Wash.) of Epidemiology*, **44**, 1388–1396 (2015).
- [183] Dhingra, R., et al., Evaluating DNA methylation age on the Illumina MethylationEPIC Bead Chip, *PLOS ONE*, **14**, e0207834 (2019).
- [184] Horvath, S., et al., The cerebellum ages slowly according to the epigenetic clock, *Ageing*, **7**, 294–306 (2015).
- [185] Handl, L., Jalali, A., Scherer, M., Eggeling, R., and Pfeifer, N., Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data, *Bioinformatics*, **35**, i154–i163 (2019).
- [186] Almén, M.S., et al., Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity, *Gene*, **548**, 61–67 (2014).
- [187] Gao, X., Thomsen, H., Zhang, Y., Breitling, L.P., and Brenner, H., The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes, *Clinical Epigenetics*, **9**, 87 (2017).
- [188] Gaunt, T.R., et al., Systematic identification of genetic influences on methylation across the human life course, *Genome Biol.*, **17**, 61 (2016).
- [189] Morrow, J.D., et al., Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-Wide association loci, *Am. J. Respir. Crit. Care Med.*, **197**, 1275–1284 (2018).
- [190] Dai, J.Y., et al., DNA methylation and cis-regulation of Gene Expr. by prostate cancer risk SNPs, *PLoS Genet.*, **16**, e1008667 (2020).
- [191] Rice, S.J., Cheung, K., Reynard, L.N., and Loughlin, J., Discovery and analysis of methylation quantitative trait loci (mQTLs) mapping to novel osteoarthritis genetic risk signals, *Osteoarthritis Cartilage*, **27**, 1545–1556 (2019).
- [192] Clark, A.D., et al., Lymphocyte DNA methylation mediates genetic risk at shared immune-mediated disease loci, *J. Allergy Clin. Immunol.*, **145**, 1438–1451 (2020).
- [193] Kim, S., et al., SNPs identified by GWAS affect asthma risk through DNA methylation and expression of cis-genes in airway epithelium, *Eur. Respir. J.*, **55**, 1902079 (2020).
- [194] Gutierrez-Arcelus, M., et al., Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing, *PLoS Genet.*, **11**, e1004958 (2015).
- [195] Kim-Hellmuth, S., et al., Cell type-specific genetic regulation of Gene Expr. across human tissues, *Science*, **369**, eaaz8528 (2020).
- [196] Peters, J.E., et al., Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease, *PLoS Genet.*, **12**, e1005908 (2016).
- [197] Shabalin, A.A., Matrix eQTL: Ultra fast eQTL analysis via large matrix operations, *Bioinformatics*, **28**, 1353–1358 (2012).
- [198] Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T., and Delaneau, O., Fast and efficient QTL mapper for thousands of molecular phenotypes, *Bioinformatics*, **32**, 1479–1485 (2016).
- [199] Pan, H., Holbrook, J.D., Karnani, N., and Kwok, C.K., Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment, *BMC Bioinf.*, **17**, 299 (2016).
- [200] Purcell, S., et al., PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses, *The American J. Hum. Genet.*, **81**, 559–575 (2007).
- [201] Scharpf, R.B., Irizarry, R.A., Ritchie, M.E., Carvalho, B., and Ruczinski, I., Using the R Package crlmm for Genotyping and Copy Number Estimation, *J. Stat. Softw.*, **40** (2011).
- [202] Ritchie, M.E., Carvalho, B.S., Hetrick, K.N., Tavaré, S., and Irizarry, R.A., R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips, *Bioinformatics*, **25**, 2621–2623 (2009).
- [203] Prive, F., Aschard, H., Ziyatdinov, A., and Blum, M.G., Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr, *Bioinformatics*, **34**, 2781–2787 (2018).
- [204] Csardi, G. and Nepusz, T., The igraph software package for complex network research, *InterJournal*, Seite 1695 (2006).
- [205] Zhu, Z., et al., Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets, *Nat. Genet.*, **48**, 481–

- 487 (2016).
- [206] Benjamini, Y. and Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300 (1995).
- [207] Lawrence, M., et al., Software for Computing and Annotating Genomic Ranges, *PLoS Comput. Biol.*, **9**, e1003118 (2013).
- [208] Hannon, E., et al., Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci, *Nat. Neurosci.*, **19**, 48–54 (2015).
- [209] Mowat, A.M. and Agace, W.W., Regional specialization within the intestinal immune system, *Nat. Rev. Immunol.*, **14**, 667–685 (2014).
- [210] Huen, K., Yousefi, P., Street, K., Eskenazi, B., and Holland, N., PON1 as a model for integration of genetic, epigenetic, and expression data on candidate susceptibility genes, *Environmental Epigenetics*, **1**, 1–11 (2015).
- [211] Volkov, P., et al., A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expr. and Metabolic Traits, *PLOS ONE*, **11**, e0157776 (2016).
- [212] Wragg, D., et al., Using regulatory variants to detect gene–gene interactions identifies networks of genes linked to cell immortalisation, *Nat. Commun.*, **11**, 343 (2020).
- [213] Jones, S., et al., Personalized genomic analyses for cancer mutation discovery and interpretation, *Sci. Transl. Med.*, **7**, 283ra53–283ra53 (2015).
- [214] Fan, Y., et al., IMAGE: high-powered detection of genetic effects on DNA methylation using integrated methylation QTL mapping and allele-specific analysis, *Genome Biol.*, **20**, 220 (2019).
- [215] Sofer, T., Schifano, E.D., Hoppin, J.A., Hou, L., and Baccarelli, A.A., A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure, *Bioinformatics*, **29**, 2884–2891 (2013).
- [216] Gatev, E., Gladish, N., Mostafavi, S., and Kober, M.S., CoMeBack: DNA methylation array data analysis for co-methylated regions, *Bioinformatics*, **36**, 2675–2683 (2020).
- [217] Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R., Merlin—rapid analysis of dense genetic maps using sparse gene flow trees, *Nat. Genet.*, **30**, 97–101 (2002).
- [218] Maurano, M.T., et al., Systematic Localization of Common Disease-Associated Variation in Regulatory DNA, *Science*, **337**, 1190–1195 (2012).
- [219] Aguirre-Gamboa, R., et al., Deconvolution of bulk blood eQTL effects into immune cell subpopulations, *BMC Bioinf.*, **21**, 243 (2020).
- [220] Lutsik, P., et al., MeDeCom: discovery and quantification of latent components of heterogeneous methylomes, *Genome Biol.*, **18**, 55 (2017).
- [221] Decamps, C., et al., Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software, *BMC Bioinf.*, **21**, 16 (2020).
- [222] Scherer, M., et al., Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using Decom-Pipeline, MeDeCom and FactorViz, *Nat. Protoc.*, **15**, 3240–3263 (2020).
- [223] Teschendorff, A.E. and Relton, C.L., Statistical and integrative system-level analysis of DNA methylation data, *Nat. Rev. Genet.*, **19**, 129–147 (2017).
- [224] Houseman, E.A., et al., DNA methylation arrays as surrogate measures of cell mixture distribution, *BMC Bioinf.*, **13**, 86 (2012).
- [225] Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S., A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies, *BMC Bioinf.*, **18**, 105 (2017).
- [226] Zheng, S.C., et al., A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix, *Epigenomics*, **10**, 925–940 (2018).
- [227] Chakravarthy, A., et al., Pan-cancer deconvolution of tumour composition using DNA methylation, *Nat. Commun.*, **9**, 3220 (2018).
- [228] Hicks, S.C. and Irizarry, R.A., methylCC: technology-independent estimation of cell type composition using differentially methylated regions, *Genome Biol.*, **20**, 261 (2019).
- [229] Salas, L.A., et al., An optimized library for reference-based deconvolution of whole-

- blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray, *Genome Biol.*, **19**, 64 (2018).
- [230] Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J., Epigenome-wide association studies without the need for cell-type composition, *Nat. Methods*, **11**, 309–311 (2014).
- [231] Rahmani, E., et al., Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies, *Nat. Methods*, **13**, 443–445 (2016).
- [232] Rahmani, E., et al., BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference, *Genome Biol.*, **19**, 141 (2018).
- [233] Houseman, E.A., et al., Reference-free deconvolution of DNA methylation data and mediation by cell composition effects, *BMC Bioinf.*, **17**, 259 (2016).
- [234] Onuchic, V., et al., Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types, *Cell Reports*, **17**, 2075–2086 (2016).
- [235] Rahmani, E., et al., Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology, *Nat. Commun.*, **10**, 3417 (2019).
- [236] Thompson, M., Chen, Z.J., Rahmani, E., and Halperin, E., CONFINED: distinguishing biological from technical sources of variation by leveraging multiple methylation datasets, *Genome Biol.*, **20**, 138 (2019).
- [237] Newman, A.M., et al., Robust enumeration of cell subsets from tissue expression profiles, *Nat. Methods*, **12**, 453–457 (2015).
- [238] Houseman, E.A., Molitor, J., and Marsit, C.J., Reference-free cell mixture adjustments in analysis of DNA methylation data, *Bioinformatics*, **30**, 1431–1439 (2014).
- [239] Everson, T.M., et al., Cadmium-Associated Differential Methylation throughout the Placental Genome: Epigenome-Wide Association Study of Two U.S. Birth Cohorts, *Environ. Health Perspect.*, **126**, 017010 (2018).
- [240] Carlström, K.E., et al., Therapeutic efficacy of dimethyl fumarate in relapsing-remitting Mult. Scler. associates with ROS pathway in monocytes, *Nat. Commun.*, **10**, 3081 (2019).
- [241] Goeppert, B., et al., Integrative Analysis Defines Distinct Prognostic Subgroups of Intrahepatic Cholangiocarcinoma, *Hepatology*, **69**, 2091–2106 (2019).
- [242] Man, Y.G., et al., Tumor-infiltrating immune cells promoting tumor Invasion Metastasis: Existing theories, *Journal of Cancer*, **4**, 84–95 (2013).
- [243] Sompairac, N., et al., Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets, *Int. J. Mol. Sci.*, **20**, 4414 (2019).
- [244] Dirkse, A., et al., Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment, *Nat. Commun.*, **10**, 1787 (2019).
- [245] Nazarov, P.V., et al., Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients, *BMC Med. Genomics*, **12**, 132 (2019).
- [246] Ritchie, M.E., et al., limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res.*, **43**, e47–e47 (2015).
- [247] Reinius, L.E., et al., Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility, *PLoS ONE*, **7**, e41361 (2012).
- [248] Jaffe, A.E. and Irizarry, R.A., Accounting for cellular heterogeneity is critical in epigenome-wide association studies, *Genome Biol.*, **15**, R31 (2014).
- [249] Therneau, T.M. and Grambsch, P.M., *Modeling Survival Data: Extending the Cox Model*, Statistics for Biology and Health, Springer New York, New York, NY (2000), ISBN 978-1-4419-3161-0.
- [250] Falcon, S. and Gentleman, R., Using GOstats to test gene lists for GO term association, *Bioinformatics*, **23**, 257–258 (2007).
- [251] The Cancer Genome Atlas Research Network, Comprehensive molecular profiling of lung adenocarcinoma, *Nature*, **511**, 543–550 (2014).
- [252] Testa, U., Castelli, G., and Pelosi, E., Lung Cancers: Molecular Characterization, Clonal Heterogeneity and Evolution, and Cancer Stem Cells, *Cancers*, **10**, 248 (2018).

- [253] Hahn, M.A., et al., Methylation of Polycomb target genes in intestinal cancer is mediated by inflammation, *Cancer Res.*, **68**, 10280 (2008).
- [254] Varambally, S., et al., The polycomb group protein EZH2 is involved in progression of prostate cancer, *Nature*, **419**, 624–629 (2002).
- [255] Cai, Y., et al., Epigenetic alterations to Polycomb targets precede malignant transition in a mouse model of breast cancer, *Sci. Rep.*, **8**, 5535 (2018).
- [256] Ward, M.J., et al., Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer, *Br. J. Cancer*, **110**, 489–500 (2014).
- [257] Travaglini, K.J., et al., A Mol. Cell atlas of the human lung from single-cell RNA sequencing, *Nature* (2020).
- [258] Robinson, M.D., McCarthy, D.J., and Smyth, G.K., edgeR: A Bioconductor package for differential expression analysis of digital Gene Expr. data, *Bioinformatics*, **26**, 139–140 (2009).
- [259] Colaprico, A., et al., TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data, *Nucleic Acids Res.*, **44**, e71 (2016).
- [260] Luo, C., et al., Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex, *Science*, **357**, 600–604 (2017).
- [261] Mulqueen, R.M., et al., Highly scalable generation of DNA methylation profiles in single cells, *Nat. Biotechnol.*, **36**, 428–431 (2018).
- [262] Lähnemann, D., et al., Eleven grand challenges in single-cell data science, *Genome Biol.*, **21**, 31 (2020).
- [263] Teschendorff, A.E., Zhu, T., Breeze, C.E., and Beck, S., EPISCOPE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data., *Genome Biol.*, **21**, 221 (2020).
- [264] Glazer, A.M., Winkelmann, R.R., Farberg, A.S., and Rigel, D.S., Analysis of Trends in US Melanoma Incidence and Mortality, *JAMA Dermatology*, **153**, 225 (2017).
- [265] Platz, A., Egyhazi, S., Ringborg, U., and Hansson, J., Human cutaneous melanoma; a review of NRAS and BRAF mutation frequencies in relation to histogenetic subclass and body site, *Molecular Oncology*, **1**, 395–405 (2008).
- [266] Luke, J.J., Flaherty, K.T., Ribas, A., and Long, G.V., Targeted agents and immunotherapies: optimizing outcomes in melanoma, *Nature Reviews Clin. Oncol.*, **14**, 463–482 (2017).
- [267] Wolchok, J.D., et al., Overall Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma, *N. Engl. J. Med.*, **377**, 1345–1356 (2017).
- [268] Scherer, M., et al., Quantitative comparison of within-sample heterogeneity scores for DNA methylation data, *Nucleic Acids Res.*, **48**, e46–e46 (2020).
- [269] Meissner, A., et al., Genome-scale DNA methylation maps of pluripotent and differentiated cells, *Nature*, **454**, 766–770 (2008).
- [270] Plongthongkum, N., Diep, D.H., and Zhang, K., Advances in the profiling of DNA modifications: cytosine methylation and beyond, *Nat. Rev. Genet.*, **15**, 647–661 (2014).
- [271] Elliott, G., et al., Intermediate DNA methylation is a conserved signature of genome regulation, *Nat. Commun.*, **6**, 6363 (2015).
- [272] Sun, Y.V., et al., Comparison of the DNA methylation profiles of human peripheral Blood Cells and transformed B-lymphocytes, *Hum. Genet.*, **127**, 651–658 (2010).
- [273] Quek, K., et al., DNA methylation intra-tumor heterogeneity in localized lung adenocarcinomas, *Oncotarget*, **8**, 21994–22002 (2017).
- [274] Wong, N.C., et al., MethPat: a tool for the analysis and visualisation of complex methylation patterns obtained by massively parallel sequencing, *BMC Bioinf.*, **17**, 98 (2016).
- [275] Lin, P., Forêt, S., Wilson, S.R., and Burden, C.J., Estimation of the methylation pattern distribution from deep sequencing data, *BMC Bioinf.*, **16**, 145 (2015).
- [276] Landau, D.A., et al., Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia, *Cancer Cell*, **26**, 813–825 (2014).
- [277] Guo, S., et al., Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA, *Nat. Genet.*, **49**, 635–642 (2017).
- [278] Landan, G., et al., Epigenetic polymorphism

- and the stochastic formation of differentially methylated regions in normal and cancerous tissues, *Nat. Genet.*, **44**, 1207–1214 (2012).
- [279] Xie, H., et al., Genome-wide quantitative assessment of variation in DNA methylation patterns., *Nucleic Acids Res.*, **39**, 4099–108 (2011).
- [280] Li, S., et al., Dynamic evolution of clonal epialleles revealed by methclone, *Genome Biol.*, **15**, 472 (2014).
- [281] Li, H., et al., The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078–2079 (2009).
- [282] Yoshihara, K., et al., Inferring tumour purity and stromal and immune cell admixture from expression data, *Nat. Commun.*, **4**, 2612 (2013).
- [283] Benelli, M., Romagnoli, D., and Demicheli, F., Tumor purity quantification by clonal DNA methylation signatures, *Bioinformatics*, **34**, 1642–1649 (2018).
- [284] Chen, X., et al., Targeting Oxidative Stress in Embryonal Rhabdomyosarcoma, *Cancer Cell*, **24**, 710–724 (2013).
- [285] Xie, W., et al., Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome, *Cell*, **148**, 816–831 (2012).
- [286] Sun, Z., Cunningham, J., Slager, S., and Kocher, J.P., Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis., *Epigenomics*, **7**, 813–828 (2015).
- [287] Kelley, D.R., Schatz, M.C., and Salzberg, S.L., Quake: quality-aware detection and correction of sequencing errors, *Genome Biol.*, **11**, R116 (2010).
- [288] Lin, P.P., Wang, Y., and Lozano, G., Mesenchymal Stem Cells and the Origin of Ewing's Sarcoma, *Sarcoma*, **2011** (2011).
- [289] Haas, S., Trumpp, A., and Milsom, M.D., Causes and Consequences of Hematopoietic Stem Cell Heterogeneity, *Cell Stem Cell*, **22**, 627–638 (2018).
- [290] Rulands, S., et al., Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency, *Cell Systems*, **7**, 63–76.e12 (2018).
- [291] Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A., and Tanay, A., Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in Cancer Cells, *Nat. Genet.*, **52**, 709–718 (2020).
- [292] Jenkinson, G., Pujadas, E., Goutsias, J., and Feinberg, A.P., Potential energy landscapes identify the information-theoretic nature of the epigenome, *Nat. Genet.*, **49**, 719–729 (2017).
- [293] Teschendorff, A.E., et al., The Dynamics of DNA Methylation Covariation Patterns in Carcinogenesis, *PLoS Comput. Biol.*, **10**, e1003709 (2014).
- [294] Park, Y. and Wu, H., Differential methylation analysis for BS-seq data under general experimental design, *Bioinformatics*, **32**, 1446–1453 (2016).
- [295] Zhang, Y., et al., PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis, *Genome Biol.*, **21**, 232 (2020).
- [296] Clark, S.J., et al., scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells, *Nat. Commun.*, **9**, 781 (2018).
- [297] Schraivogel, D., et al., Targeted Perturb-seq enables genome-scale genetic screens in single cells, *Nat. Methods*, **17**, 629–635 (2020).
- [298] Cao, K., Bai, X., Hong, Y., and Wan, L., Unsupervised topological alignment for single-cell multi-omics integration, *Bioinformatics*, **36**, i48–i56 (2020).
- [299] Argelaguet, R., et al., MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data, *Genome Biol.*, **21**, 111 (2020).
- [300] Philpott, M., Cribbs, A.P., Brown, T., Brown, T., and Oppermann, U., Advances and challenges in epigenomic single-cell sequencing applications, *Curr. Opin. Chem. Biol.*, **57**, 17–26 (2020).
- [301] Vaisvila, R., et al., EM-seq: Detection of DNA Methylation at Single Base Resolution from Picograms of DNA, *bioRxiv* (2020).
- [302] Wierzbinska, J.A., et al., Methylome-based cell-of-origin modeling (Methyl-COOM) identifies aberrant expression of immune regulatory molecules in CLL, *Genome Med.*, **12**, 29 (2020).
- [303] Akulenko, R., Merl, M., and Helms, V., BEclear: Batch Effect Detection and Adjustment

- in DNA Methylation Data, *PLOS ONE*, **11**, e0159921 (2016).
- [304] Hebestreit, K., Dugas, M., and Klein, H.U., Detection of significantly differentially methylated regions in targeted bisulfite sequencing data, *Bioinformatics*, **29**, 1647–1653 (2013).
- [305] Warden, C.D., et al., COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis, *Nucleic Acids Res.*, **41**, e117–e117 (2013).
- [306] Catoni, M., Tsang, J.M., Greco, A.P., and Zabet, N.R., DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts, *Nucleic Acids Res.*, **46**, e114 (2018).
- [307] Peters, T.J., et al., De novo identification of differentially methylated regions in the human genome, *Epigenetics Chromatin*, **8**, 6 (2015).
- [308] Xu, Z., Niu, L., Li, L., and Taylor, J.A., ENmix: a novel background correction method for Illumina HumanMethylation450 Bead-
Chip, *Nucleic Acids Res.*, **44**, e20–e20 (2016).
- [309] van Iterson, M., et al., MethylAid: visual and interactive quality control of large Illumina 450k datasets, *Bioinformatics*, **30**, 3435–3437 (2014).
- [310] Kishore, K., et al., methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data, *BMC Bioinf.*, **16**, 313 (2015).
- [311] Jühling, F., et al., metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data, *Genome Res.*, **26**, 256–262 (2016).
- [312] Phipson, B., Maksimovic, J., and Oshlack, A., missMethyl: an R package for analyzing data from Illumina’s HumanMethylation450 platform, *Bioinformatics*, **32**, 286–288 (2016).
- [313] Fortin, J.P., Fertig, E., and Hansen, K., shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R, *F1000Research*, **3**, 175 (2014).
- [314] Müller, F., Analyzing DNA Methylation Signatures of Cell Identity, Ph.D. thesis, Saarland University (2016).